BMJ Open

Accuracy of consumer-grade smart devices to measure sleep compared with polysomnography, in a sleep disorders population

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-044015
Article Type:	Original research
Date Submitted by the Author:	26-Aug-2020
Complete List of Authors:	Ellender, Claire; Princess Alexandra Hospital, ; The University of Queensland, Meaklim, Hailey; Melbourne Sleep Disorders Centre; St Vincent's Hospital Melbourne Pty Ltd Joyce, Rosemarie; Melbourne Sleep Disorders Centre Cunnington, David; Melbourne Sleep Disorders Centre, Swieca, John; Melbourne Sleep Disorders Centre
Keywords:	SLEEP MEDICINE, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, RESPIRATORY MEDICINE (see Thoracic Medicine)

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ORIGINAL RESEARCH

Accuracy of consumer-grade smart devices to measure sleep compared with polysomnography, in a sleep disorders population

Claire M. Ellender[1,2]

Rosemarie Joyce[1]

Hailey Meaklim[1]

David Cunnington[1]

John Swieca[1]

[1] Melbourne Sleep Disorders Centre, East Melbourne, Victoria, Australia

[2] Princess Alexandra Hospital, Brisbane, Australia

ORCID's: Hailey Meaklim 0000-0003-0448-3567; Claire Ellender 0000-0002-1727-576X; David Cunnington 0000-0002-8403-0420; John Swieca 0000-0001-8281-4048.

Correspondence: Dr John Swieca, Melbourne Sleep Disorders Centre

Suite 508, 100 Victoria Parade, East Melbourne, VIC 3002 Australia.

Tel +61 3 9663 1993, Fax +61 3 9663 1553

Email swieca@msdc.com.au

Word Count: 2554

Keywords: Consumer sleep monitor; accelerometer, polysomnography; sleep; sleep measurement; validation; wearables

ABSTRACT:

Objectives: Consumer-grade smart devices are increasingly being used to measure waking activity and sleep; however the ability of these devices to accurately measure sleep in clinical populations warrants more examination. The aim of the present study was to assess the accuracy of consumer-grade sleep monitors to through comparison with gold standard polysomnography (PSG).

Design: A prospective cross-sectional cohort study was performed.

Setting: Adults undergoing PSG for investigation of a suspected sleep disorder.

Participants: 54 sleep-clinic patients were assessed using three consumer-grade sleep monitors (Jawbone UP3®, ResMed S+® and Beddit®) in addition to PSG.

Outcomes: Jawbone UP3®, ResMed S+® and Beddit® were compared with gold standard in-laboratory polysomnography on 4 major sleep parameters - total sleep time (TST), sleep onset latency (SOL), Wake After Sleep Onset (WASO) and sleep efficiency (SE).

Results: The accelerometer Jawbone UP3® was found to overestimate TST by 28mins, with reasonable reliability compared with gold standard for TST, WASO and SE. The doppler radar ResMed S+® device was found to underestimate TST by 34mins however had poor reliability compared with PSG for TST, SOL and SE. The mattress device, Beddit®, had the least reliability; underestimated TST by 53mins on average and poor reliability compared to PSG for all measures. High device synchronisation failure occurred, with 20% of recordings incomplete due to Bluetooth drop out and recording loss.

Conclusion: The Jawbone UP3® had the strongest reliability in sleep measurements compared with PSG. Consumer grade devices assessed do not have strong enough agreement with gold standard measurement to replace clinical evaluation and PSG sleep testing, however are an opportunity as powerful patient engagement tools for long-term sleep measurement.

Strengths and Limitations of this study

- Consumer grade devices were compared with gold standard in clinic patients.
- More than one device was included for comparison.
- This study includes measure of sleep parameters that clinicians frequently need to review in daily practice, such as total sleep time and sleep efficiency.
- High device failure was found in this study, confirming that consumer grade devices cannot be used to replace high fidelity diagnostic measurement.
- This sample had patients with sleep apnoea, insomnia or hypersomnia as their final sleep diagnosis.

Patient and Public involvement

Patients at our sleep disorders centre sparked the initial interest into assessing the accuracy of consumer-grade sleep monitors. Our clinicians were often asked how about the accuracy of home sleep monitors. To answer this question our team invited the patients of our clinic to be involved in evaluating three commonly available consumer grade devices. Participants were not paid for their involvement but did provide written consent. The findings of this research suggest that consumer-grade sleep monitors can give insights into trends in sleep, but are not accurate enough to replace laboratory measurement.

BACKGROUND

Poor sleep quality and duration has been shown to be an independent risk to overall mortality and for many chronic diseases.[1] The gold standard test for the measurement of sleep and diagnosis of sleep disorders is attended polysomnography (PSG). However, this is an involved and costly test that requires complex equipment, dedicated space, trained staff, and does not lend itself well to multi-night monitoring.

Sales of consumer sleep monitors and wearable consumer-grade smart devices have dramatically increased in recent years, with 33 million units estimated to have been sold in the United States in 2015[2] and the estimated value of the wearable industry in the USA expected to grow to 8.5 billion in 2020.[3,4] Consumer-grade devices fall into three major categories (i) wrist based devices (eg Jawbone, FitBit); (ii) Bedside devices (eg ResMed S+®, Touch-Free Life Care®); and (iii) Mattress-based devices (eg Beddit®, EarlySense Mattress®, Emfit Bed Sensor®). Each of the categories of devices utilise unique proprietary algorithms for inferring wake/sleep, body position and measures of sleep quality.

The Jawbone UP® (the precursor to the UP3 used in this study) has been compared to PSG in adolescents and concluded to have good agreements for TST, SE and wake after sleep onset (WASO), however the tendency to underestimate TST and sleep efficiency increased with age.[5] In a study of adult women, the FitBitChargeHR® overestimated TST by 27min, and was found to have significantly different SOL and WASO compared to PSG.[5] Similarly in adolescents the Jawbone UP® tended to overestimate TST and SOL, whilst underestimating WASO. The researchers also found greater discrepancies in nights when participants had more disrupted sleep (ie lower TST and greater SOL and WASO).[6]

In patients with suspected central disorders of hypersomnolence, the Jawbone UP3® was found to significantly overestimate TST by an average of 39.6 minutes compared to PSG and was not able to discriminate stages of sleep adequately.[7] Interestingly, the Jawbone UP3® performed similarly to actigraphy in this study. Another clinical study found that the FitBit Flex® overestimated TST more in a group of insomnia patients compared to good sleepers (32.9 mins vs. 6.5 mins).[8] Taken together, these

two studies suggest that consumer-grade sleep devices are less accurate for TST measurement clinical sleep disorder populations than they are for good sleepers.

Beddit in 10 health controls was found to have poor agreement with TST, WASO and SE.[9] SOL was the only measure to have agreement, but had a wide variance.[9] The sensor technology used in the ResMed S+® been shown to have moderate accuracy in measuring TST and sleep efficiency in healthy volunteers compared to PSG.[10] Furthermore its utility in measuring sleep disordered breathing has been investigated and found to have reasonable accuracy in detecting moderate obstructive sleep apnoea, with a sensitivity of 89% and specificity 92%.[11]

Patients are increasingly attending sleep clinics with downloads from these devices for discussion with primary care physicians and sleep specialists, or asking clinicians which consumer grade device is best to track sleep at home. These commonly encountered situations in the sleep clinic raise the questions: how reliable are consumer grade devices and which type of technology is most accurate compared to gold standard? This study aims to answer these questions with an in-laboratory comparison of PSG with the three consumer devices - Jaw Bone UP3®, Beddit® and ResMed S+® in a sleep clinic population. It was hypothesized that these devices would have similar accuracy in detecting TST, SOL, WASO and SE.

METHODS

Study Population

54 adult patients were consecutively recruited through a private sleep disorders centre in Melbourne, Australia from June 2015 to February 2016. Inclusion criteria were age >18years and any patient who required overnight polysomnography as standard investigation following sleep physician review to either confirm or exclude sleep disordered breathing. All patients attending the laboratory for a polysomnogram were screened for inclusion. Exclusion criteria were age <18years, positive airway pressure titration study, pregnancy and cognitive impairment.

Procedure

All assessments took place at an attended sleep laboratory in Melbourne, Australia. The study was approved by the Human Research and Ethics Committee of St

Vincent's Hospital, Melbourne (LRR141/15). Sleep laboratory staff were trained to set up the 3 devices in addition to regular overnight polysomnography monitoring. The primary outcome measure was Total Sleep Time (TST) and secondary outcomes were sleep onset latency (SOL, min), sleep efficiency (SE, %) as TST/(TST+ total wake time) and wake after sleep onset (WASO, min). Figure 1 demonstrates the consort statement.

Polysomnographic Recording

PSG was measured using a standard six-channel electroencephalography, submental electromyography and electrooculography, electrocardiogram, airflow (thermistor & nasal cannula), respiratory effort, oximetry, snoring (dB sound meter), body position, pulse rate, leg electromyography and digital video, recorded according to American Academy of Sleep Medicine standards.[12] The following standard sleep parameters were recorded via PSG: Total sleep time (TST), sleep onset latency (SOL, min), total wake time (TWT, min), sleep efficiency (SE, %) as TST/(TST+TWT) and wake after sleep onset (WASO, min). Participants were classified as having obstructive sleep apnoea if the apnoea hypopnoea index was >5 events/hr. The scientist scoring the PSG was blinded to the download of consumer grade devices.

JawBone UP3®

Participants were fitted on the participant's non dominant wrist with the Jawbone UP3® shortly before lights out time. Data was collected via a dedicated iPod Touch, synced to the Jawbone® app.[13] This consumer-grade actigraphy device has a three-axis accelerometer and heart rate monitor, which together measure TST, SOL, WASO and SE which were exported by a technician the following morning after the PSG was complete.

ResMed S+®

The ResMed S+® is a non-contact radio-frequency sensor that continuously measures the biomotion due to breathing and body-movement of the participant subject in bed. The sensor operates in a license-free band at 5.8 GHz, emits an average power less than 1 mV and is capable of sensing movement and breathing over a distance ranging from 0.3 to 1.5 meters. The device was positioned by the bedside and synced

shortly before lights out time to a dedicated iPod with the ResMed S+® app. [14] Measurements from the ResMed S+® were TST, SOL, WASO, SE which were exported by a technician the following morning after the PSG was complete.

Beddit®

The primary sensor in the Beddit® is a piezoelectric 70cm band that was attached to the mattress prior to patients getting into bed. The device detects micro-movements of the chest wall from heartbeats and respiration and uses ballistocardiography to infer sleep stage and time. Ballistocardiography is a non-invasive measurement of cardiac output and respiration by converting mechanical motion (e.g. movement generated by a heartbeat) to a digital signal. Measurements from the Beddit® were taken each night using the device synced to a dedicated iPod running the Beddit® app.[15] Output from the app included TST, SOL, WASO, SE and HR which were exported by a technician the following morning after the PSG was complete.

Statistical analyses

Each of the three non-invasive devices was compared with PSG as the gold standard on an intention to treat basis. The primary and secondary outcomes were compared on total measurements over the night, not epoch-by-epoch method. Normality was assessed using the Shapiro-Wilk's test. Agreement between gold standard and test device, were assessed for reliability using Intraclass Correlation Coefficients (ICCs) with two-way Random-Effects model and Bland-Altman plots[16]. Reliability was considered acceptable if the ICC was greater than 0.5, reliability was considered 'good' if ICC was 0.7-0.9 and 'excellent' if >0.9.

RESULTS

Fifty four adult patients (31 females, 23 males; mean age 54 years ±SD 18years) participated, see Table 1 for patient demographics. The final sleep diagnosis found was obstructive sleep apnoea in 33 (61%), insomnia 9 (17%) and central hyper-somnolence disorder in 12 (22%).The mean PSG detected TST was 371min (SD±69), SOL of 16min (SD±15), WASO 63min (SD±56) and SE of 82% (SD±13%). The absolute values of the

measurements for each device is summarised in Table 2, and the mean differences and intra-class correlation in Table 3.

The Jawbone UP3® was found to have overestimated TST by 28 min (95% CI: -155-98). The intra-class correlation coefficient between the PSG TST and Jawbone was 0.6 (95% CI: 0.34-0.77), indicating 'good' reliability between the two tests. 4 out of the 42 data points fell outside the 95% confidence interval, with greatest reliability seen when TST was between 300-400min. SOL was over estimated by Jawbone UP3® by 0.14 min (95% CI: -39 – 39), with an intra-class correlation was 0.29 (95% CI:-0.04 – 0.57) indicating poor to moderate reliability between the methods. Only 1 data point fell outside of the 95% CI, with negative bias present for measurements greater than 15 minutes. WASO was overestimated using Jawbone UP3® by only 1.7min (95% CI: -103 – 100), with 3 of the data points falling outside of the 95% CI and negative bias when measurements were over 50 minutes. The ICC between by Jawbone UP3® and PSG for WASO 0.55 (95% CI: 0.29-0.73), indicating 'moderate' agreement. Sleep efficiency was underestimated by 0.5 min (95% CI: -18–20), with bias found with measurements less than 85%. The sleep efficiency ICC for Jawbone Up3® was 0.65 indicating 'good' reliability between the two measures.

The ResMed S+® underestimated TST by 34 min (95% CI:-183-252), with 3 measurements outside the 95% CI, and ICC of 0.36 (95% CI: 0.02-0.63) indicating 'poor' reliability. SOL was poorly measured by the ResMed S+® with 35min (95% CI: -127-55) overestimation and ICC -0.01 (95% CI:-0.21 – 0.26) and negative bias with measurements greater than 30min. WASO however was reliably measured by ResMed S+® with overestimation 27min (95% CI: -125-71), ICC 0.61 (95% CI: 0.28 – 0.8). Sleep efficiency was not reliably measured with the ResMed S+®, with underestimation by 15min (95% CI:-21–53) and ICC only 0.06 (95% CI: -0.06-0.58).

The Beddit® and PSG had the least agreement, with ICC all <0.5 indicating 'poor' reliability. TST was underestimated by 53min (95% CI:-128 - 235), ICC 0.4 (95% CI: 0.09 – 0.63) and 4 measurements outside the 95th CI. SOL overestimated by 44min (95% CI: -160-71), ICC 0.01 (95% CI:-.173-0.2). SE 1.35min (95% CI:-35-38) and ICC 0.05 (95% CI:-0.04-0.51).

Figure 2 demonstrates agreement for the three devices compared with PSG displayed as Blant-Altman Plots and Table 3 summarises the mean differences and intra-class correlation coefficients. No significant were found between PSG and the test devices were mediated by gender, age or final sleep diagnosis for any of the measured parameters (p-values all >0.05).

Consumer-grade recording failure

Consumer-grade devices were set-up by Sleep Scientist staff each night at the time of the standard PSG set-up. Despite this, device or recording failure resulting in inability to record sufficient data, on the single night of recording, in the consumer-grade devices was common. Failure to synchronise with the dedicated Bluetooth device was the most common reason for device failure. The ResMed S+® failed to synchronise the most, with 25/54 nights (46%) resulting in recording failure. The Jawbone and Beddit had similar rates of synchronisation failure (12/54, 22%), however not usually in the same room or on the same patient. Comparisons were made on an intention to treat analysis, even where large differences in TST were seen.

DISCUSSION

The accuracy of these three consumer-grade smart devices has simultaneously been compared with gold standard attended PSG in an adult sleep clinic cohort. For each of the devices, there were components of sleep measurement with significant agreement to the gold standard. In regards to the primary outcome measure of TST, only the Jawbone UP3® had strong agreement with PSG, with an overestimation of 28min. The Jawbone UP3® had the best agreement across secondary outcomes of WASO and SE. The Beddit® had the least agreement with PSG, all components having poor reliability when compared with gold standard PSG.

Wearable devices, particularly wrist-worn accelerometers have now been widely compared with PSG. Similar to the results of this study, the accelerometers have been shown to overestimate total sleep time by around 20-30minutes, particularly in sleep disordered populations compared with healthy controls.[5,6,8] Previous investigations into consumer grade accelerometers in clinical populations found TST overestimated by

32.9min[8] in a population of 33 insomnia patients and 39min in 43 hyper-somnolence patients[7]. In our study SOL had a large confidence interval, with bias found with measurements over 15min, consistent with findings of a recent systematic review and meta-analysis.[17]

The Beddit® device is one of the least reviewed consumer grade devices out of the three assessed in this study. Tuominen *et al.* (2019) found in 10 healthy controls the Beddit overestimated total sleep time by 43min, whereas our data suggests a significant underestimation (PSG TST 371min versus Beddit® TST 321 min) with a larger sample size (n = 42). Tuominen *et al.* (2019) was also able to access WASO data, which was not available with the model of Beddit® tested in this study, and found to underestimate WASO by 32min. Non-wearable devices have a potential growing market as non-intrusive home monitors of sleep, as they can be applied in a "set and forget" method. Thus further refinement and evaluation of bed-based devices would be desirable.

The high device synchronisation failure rate in our study is concerning, despite the set-up being performed by sleep laboratory scientific staff. There is no way to calibrate these consumer-grade devices over time and it is difficult to monitor device connectivity to the Bluetooth device until the next morning. Moving forward, these finding should indicate to developers, that some data storage is needed within sleep monitors to mitigate synchronisation failure. The high failure rate further confirms the role of these consumer devices is not to replace that of a diagnostic sleep study.

The main strength of this study was the sample size and that it was conducted in a clinical adult sleep population with a range of suspected sleep disorders. This makes the findings more translatable to clinicians managing patients with sleep disorders. Further, assessing a number of different devices is a novel approach. The weaknesses of the study include a high device recording failure rate, predominantly with Bluetooth synchronisation failure. A further weakness was that actigraphy was not directly compared with the consumer grade devices.

This study indicates that the wrist worn Jawbone UP3® had the best reliability in measuring sleep compared with gold standard and can provide useful information about commonly measured parameters of sleep quality. For Sleep Medicine Clinicians, the

translation of these findings, is that when our patients present with longitudinal measurements of sleep from their consumer grade devices, we can be reassured that wrist worn devices have reasonably reliability and can be harnessed as an engagement tool for behavioural sleep interventions. This is consistent message with the American Academy of Sleep Medicine's position statement about the use of consumer-grade sleep devices stating that these devices cannot be used for clinical diagnosis, however they allow for meaningful discussions with patients about sleep and encourage active participation in sleep-related health care.[18]

CONCLUSION

Given the large body of literature linking sleep quality to mortality and many chronic diseases, patient-collected longitudinal sleep data provides a powerful insight into a patient's overall health. This study adds to the data of consumer grade wearable sleep monitors, showing they can provide some reliable information compared to gold standard PSG, however do not replace clinical evaluation and gold-standard PSG sleep testing. In reviewing sleep data collected by patients with consumer-grade devices, clinicians are encouraging measurement and quantification of sleep, which in turn will likely emphasise the importance of quality sleep in maintaining good health.

ACKNOWLEDGMENTS

Sleep laboratory staff at St Vincent's Private Hospital, East Melbourne for their set up efforts. Telstra Corporation Ltd (Australia) for the provisions of the Jawbone UP3, ResMed (San Diego) for the ResMed S+ and Beddit Ltd (Finland) for the supply of the test devices used. Acknowledgement to Dr Farah Zahir, Qfab, Queensland Cyber Infrastructure Foundation for bio-statistical support and interpretation of the data presented.

Dataset availability

The dataset will be available upon emailed request to the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public or not-for-profit sectors.

Competing Interests

The Telstra Corporation Ltd (Australia) provided the Jawbone UP3 test devices used in the study, ResMed (San Diego) provided the ResMed S+ and Beddit Ltd (Finland) provided the Beddit device. The data reported in this manuscript was presented as a poster at the 27th Annual Scientific Meeting of the Australasian Sleep Association and the Australasian Sleep Technologists Association, held on 22–24 October 2015, in Melbourne, Australia.

Author contributions

Claire M. Ellender	Protocol preparation, Participant consent, Data collection, Data analysis, Manuscript preparation
Rosemarie Joyce	Participant consent, Data collection, Manuscript preparation
Hailey Meaklim	Data analysis, Manuscript preparation
David Cunnington	Protocol preparation, Data analysis, Manuscript preparation
John Swieca	Protocol preparation, Data analysis, Manuscript preparation

Table 1 Patient demographics

Variable	Results (n = 54)
Age in years, mean (SD)	54 (SD±18)
Gender	31 (57%) women
	23 (43%) men
BMI kg/m², median (IQR)	27 (24-31)
PSG AHI events/hr, median (IQR)	9 (3-18.75)
Indication for PSG	
Rule in suspected OSA	32 (60%)
Rule out OSA	22 (40%)
Final clinical diagnosis	
OSA syndrome	33 (61%)
Insomnia	9 (17%)
Hypersomnia	12 (22%)

PSG, Polysomnogram; BMI, Body Mass Index; AHI, Apnoea hypopnoea index; OSA, Obstructive sleep apnoea

Table 2 Mean sleep duration

	DEVICE			
VARIABLE	PSG	Jawbone UP3® (N = 42)	ResMed S+® (N = 29)	Beddit® (N = 42)
TST (MIN SD±)	371 ±69	397 ±83	345.8 ±120	321 ±107
SOL (MIN)	16 ±15	18 ±16	50 ±44	60 ±57
WASO (MIN)	63 ±56	65 ±55	80 ±72	-
SE (%)	82.4 ±13	82.9 ±11	68.8 ±21	81 ±17

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency.

Table 3 Mean difference in test device and PSG

	MEAN DIFFERENCES & INTRA-CLASS CORRELATION COEFFICIENT BETWEEN PSG AND TEST DEVICE		
VARIABLE	PSG vs Jawbone UP3®	PSG vs ResMed S+®	PSG vs Beddit®
TST (MIN)	-28.57 (95% CI:-155-98)	+34 (95% CI: -183–252)	+53 (95% CI:-128 – 235)
ICC	0.6*	0.36	0.01
SOL (MIN)	-0.14 (95% CI:-39–39)	-35.6 (95% CI:-127–55)	-44.6 (95% CI:-160–71)
ICC	0.29	-0.01	0.04
WASO (MIN)	-1.7 (95% CI:-103–100)	-27 (95% CI:-125–71)	-
ICC	0.55*	0.61	-
SE (%)	+0.5min (95% CI:-18–20)	+15 (95% CI:-21–53)	-1.35 (95% CI:-35–38)
ICC	0.65*	0.06	0.26

An overestimation is expressed as a negative (-) mean difference and an underestimation is expressed as a positive (+) mean difference value.

ICC >0.5 is marked as * and shaded as moderate reliability between PSG and test measure

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency.

References

1. Cai H, Shu XO, Xiang YB, et al. Sleep duration and mortality: a prospective study of 113 138 middle-aged and elderly Chinese men and women. Sleep 2015;38:529-36.

2. Davona T. The Wearables Report: Growth Trends, Consumer Attitudes and Why Smart Watches Will Dominate, Business Insider, February 12, 2015. Buisness Insider Australia. Sydney: Allure Media; 2015.

3. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. Physiology & behavior 2016;158:143-9.

4. US Enterprise Wearables Market: 5-Year Forecast, 2014–2019. 2015. at http://www.marketwired.com/press-release/compass-intelligence-forecasts-wearables-enterprise-grow-exponentially-us-device-revenue-2032309.htm.)

5. de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults. Chronobiol Int 2015;32:1024-8.

6. de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents. Sleep 2015;38:1461-8.

7. Cook JD, Prairie ML, Plante DT. Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy. J Clin Sleep Med 2018;14:841-8.

8. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. J Psychosom Res 2017;97:38-44.

9. Tuominen J, Peltola K, Saaresranta T, Valli K. Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. J Clin Sleep Med 2019;15:483-7.

10. De Chazal P, Fox N, O'Hare E, et al. Sleep/wake measurement using a non-contact biomotion sensor. J Sleep Res 2011;20:356-66.

11. Zaffaroni A, de Chazal P, Heneghan C, Boyle P, Mppm PR, McNicholas WT. SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea

index. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2009;2009:7091-4.

12. Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. J Clin Sleep Med 2012;8:597-619.

13. Jawbone UP3. 2016. (Accessed 19/4/16, 2016, at https://jawbone.com/support/articles/000001027/download-the-app.)

14. 2016. (Accessed 19/4/16, 2016, at https://itunes.apple.com/us/app/s+-by-resmed/id883611019?mt=8.)

15. Beddit Sleep Tracker. 2016. (Accessed 19/4/16, 2016, at http://support.beddit.com/hc/en-us/articles/201422237-Downloading-the-Beddit-app.)

16. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet 1986;1:307-10.

17. Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. Sleep Med Rev 2020;49:101227.

18. Khosla S, Deak MC, Gault D, et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. J Clin Sleep Med 2018;14:877-80.

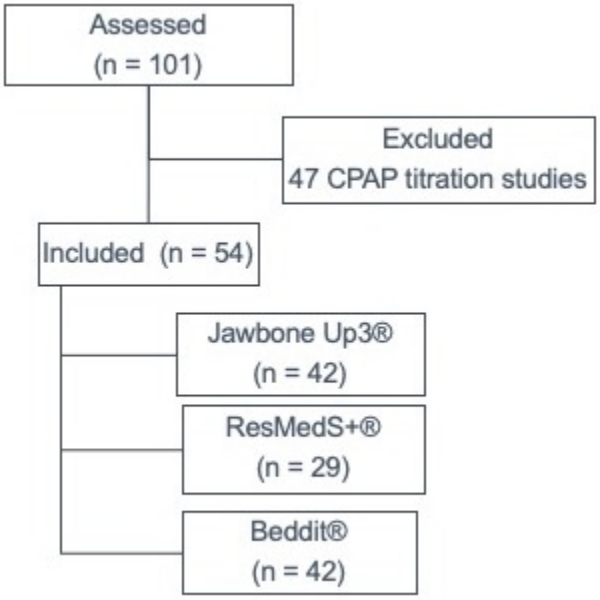


Figure 1 - Consort statement of data collection.

156x153mm (54 x 54 DPI)

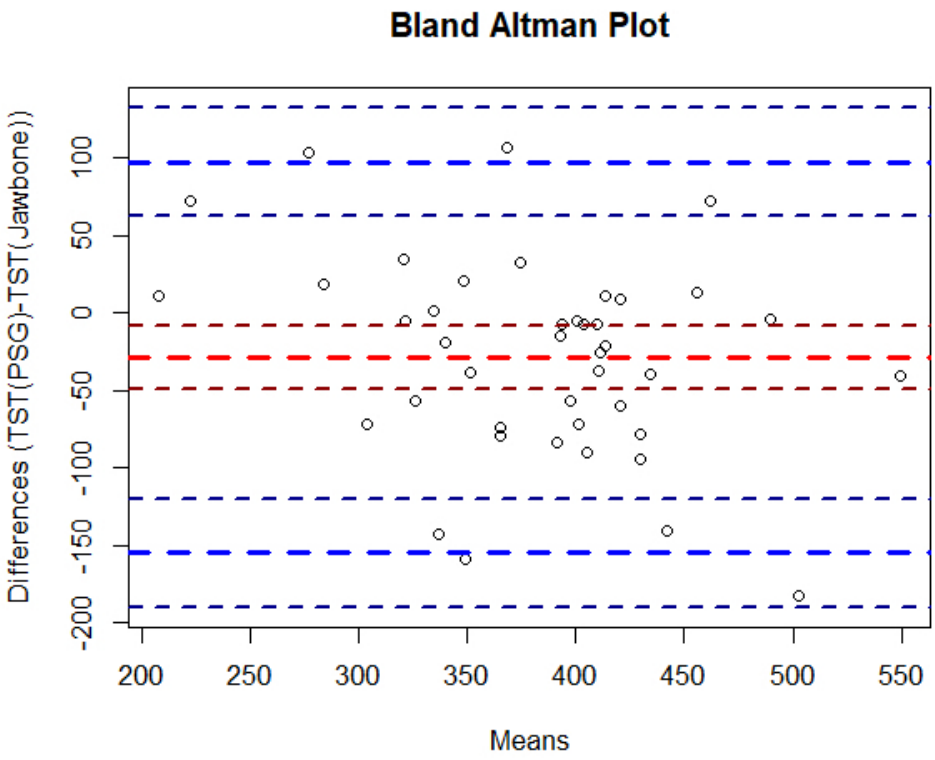


Figure 2(a) PSG vs Jawbone UP3
Total Sleep Time (TST) agreement between Polysomnography (PSG) and test devices. The mean difference is shown in a red long dashed line (with 95% confidence intervals), and 95th upper and lower confidence limits in long blue dash lines (with 95% confidence intervals).

152x127mm (100 x 100 DPI)

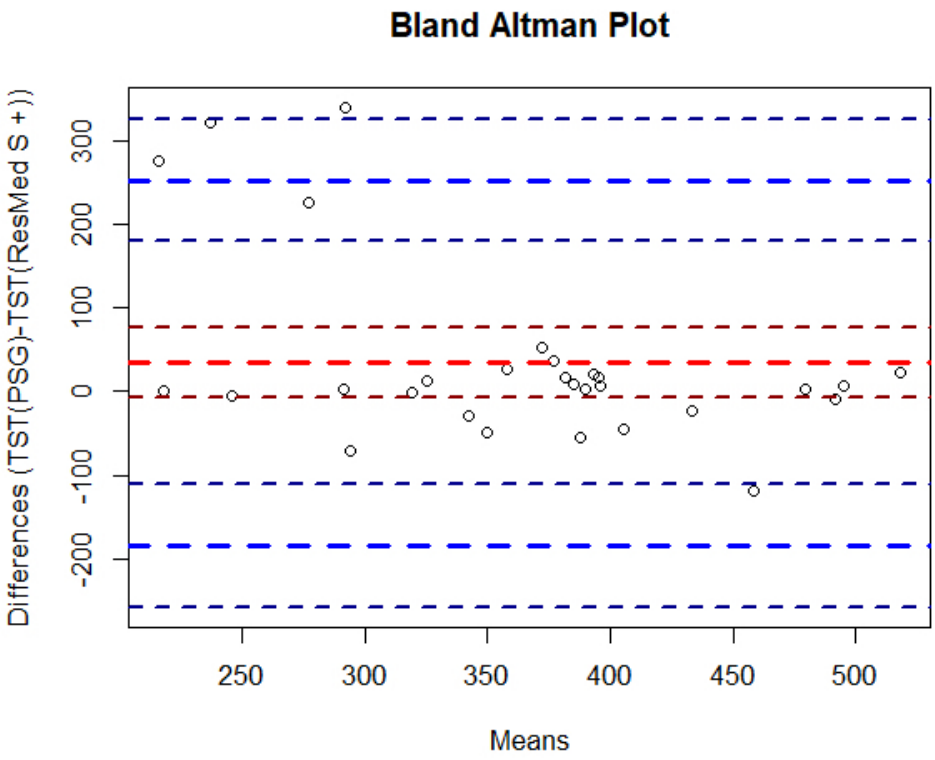


Figure 2(b) PSG vs ResMed S+
Total Sleep Time (TST) agreement between Polysomnography (PSG) and test devices. The mean difference is shown in a red long dashed line (with 95% confidence intervals), and 95th upper and lower confidence limits in long blue dash lines (with 95% confidence intervals).

152x127mm (100 x 100 DPI)

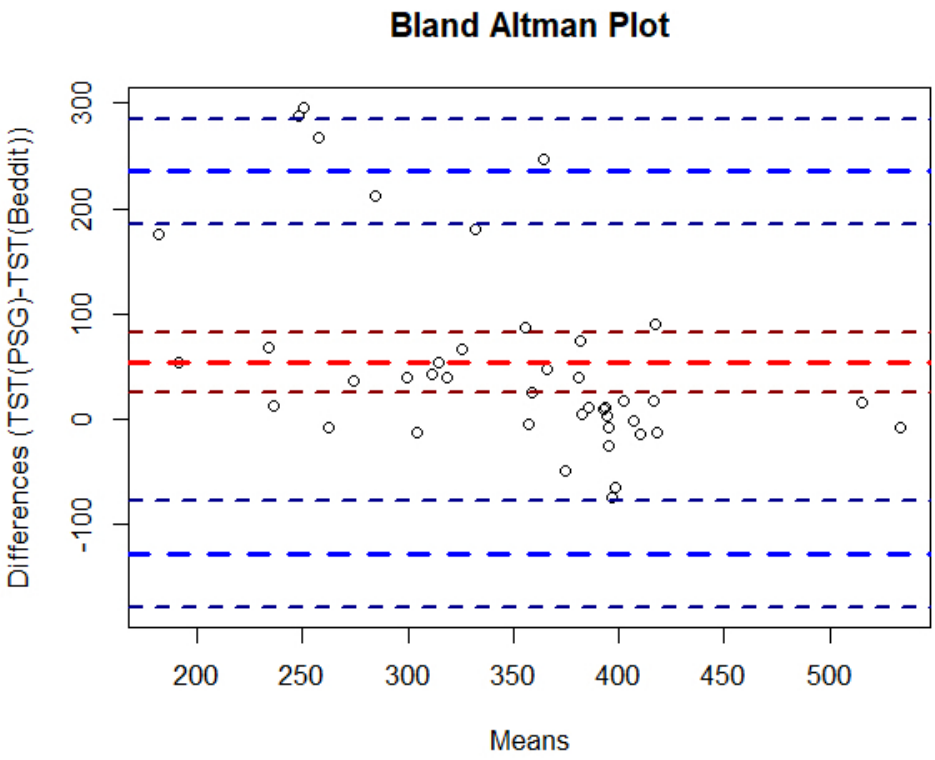


Figure 2(c) PSG vs Beddit
Total Sleep Time (TST) agreement between Polysomnography (PSG) and test devices. The mean difference is shown in a red long dashed line (with 95% confidence intervals), and 95th upper and lower confidence limits in long blue dash lines (with 95% confidence intervals).

152x127mm (100 x 100 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3
	4	Study objectives and hypotheses	3
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	4
Participants	**6**	Eligibility criteria	4
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	4
	8	Where and when potentially eligible participants were identified (setting, location and dates)	4
	9	Whether participants formed a consecutive, random or convenience series	4
Test methods	**10a**	Index test, in sufficient detail to allow replication	4
	10b	Reference standard, in sufficient detail to allow replication	4
	11	Rationale for choosing the reference standard (if alternatives exist)	4
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	-
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	4
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	4
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	4
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	4-5
	15	How indeterminate index test or reference standard results were handled	5
	16	How missing data on the index test and reference standard were handled	5
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	5
	18	Intended sample size and how it was determined	-
RESULTS			
Participants	**19**	Flow of participants, using a diagram	9
	20	Baseline demographic and clinical characteristics of participants	8
	21a	Distribution of severity of disease in those with the target condition	8
	21b	Distribution of alternative diagnoses in those without the target condition	8
	22	Time interval and any clinical interventions between index test and reference standard	-
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	9
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	9
	25	Any adverse events from performing the index test or the reference standard	-
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	7
	27	Implications for practice, including the intended use and clinical role of the index test	8
OTHER INFORMATION			
	28	Registration number and name of registry	
	29	Where the full study protocol can be accessed	
	30	Sources of funding and other support; role of funders	

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

Accuracy of consumer-grade smart devices to measure sleep compared with polysomnography, in a sleep disorders population

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-044015.R1
Article Type:	Original research
Date Submitted by the Author:	11-May-2021
Complete List of Authors:	Ellender, Claire; Princess Alexandra Hospital, ; The University of Queensland, Zahir, Syeda ; The University of Queensland, QCIF Facility for Advanced Bioinformatics Meaklim, Hailey; Melbourne Sleep Disorders Centre; St Vincent's Hospital Melbourne Pty Ltd Joyce, Rosemarie; Melbourne Sleep Disorders Centre Cunnington, David; Melbourne Sleep Disorders Centre, Swieca, John; Melbourne Sleep Disorders Centre
Primary Subject Heading:	Respiratory medicine
Secondary Subject Heading:	General practice / Family practice
Keywords:	SLEEP MEDICINE, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, RESPIRATORY MEDICINE (see Thoracic Medicine), Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ORIGINAL RESEARCH

Accuracy of consumer-grade smart devices to measure sleep compared with polysomnography, in a sleep disorders population

Claire M. Ellender[1,2]

Syeda Farah Zahir[3]

Hailey Meaklim[1]

Rosemarie Joyce[1]

David Cunnington[1]

John Swieca[1]

[1] Melbourne Sleep Disorders Centre, East Melbourne, Victoria, Australia

[2] Princess Alexandra Hospital, Brisbane, Australia

[3] QCIF Facility for Advanced Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, 4072, AUSTRALIA

ORCID's: Hailey Meaklim 0000-0003-0448-3567; Claire Ellender 0000-0002-1727-576X; David Cunnington 0000-0002-8403-0420; Syeda Farah Zahir 0000-0002-2074-6999; John Swieca 0000-0001-8281-4048.

Correspondence: Dr Claire Ellender

Department of Respiratory & Sleep Medicine, Princess Alexandra Hospital, 199 Ipswich Rd, Brisbane, Qld, Australia 4170

Tel +61731762698, Fax +61 731766170

Email Claire.ellender@health.qld.gov.au

Word Count: 3180

Keywords: Consumer sleep monitor; accelerometer, polysomnography; sleep; sleep measurement; validation; wearables

ABSTRACT:

Objectives: Consumer-grade smart devices are increasingly being used to measure waking activity and sleep; however the ability of these devices to accurately measure sleep in clinical populations warrants more examination. The aim of the present study was to assess the accuracy of consumer-grade sleep monitors compared with gold standard polysomnography (PSG).

Design: A prospective cohort study was performed.

Setting: Adults undergoing PSG for investigation of a suspected sleep disorder.

Participants: 54 sleep-clinic patients were assessed using three consumer-grade sleep monitors (Jawbone UP3®, ResMed S+® and Beddit®) in addition to PSG.

Outcomes: Jawbone UP3®, ResMed S+® and Beddit® were compared with gold standard in-laboratory polysomnography on 4 major sleep parameters - total sleep time (TST), sleep onset latency (SOL), Wake After Sleep Onset (WASO) and sleep efficiency (SE).

Results: The accelerometer Jawbone UP3® was found to overestimate TST by 28mins, with reasonable agreement compared with gold standard for TST, WASO and SE. The doppler radar ResMed S+® device underestimated TST by 34mins however had poor absolute agreement compared with PSG for TST, SOL and SE. The mattress device, Beddit® underestimated TST by 53mins on average and poor reliability compared to PSG for all measures except TST. High device synchronisation failure occurred, with 20% of recordings incomplete due to Bluetooth drop out and recording loss.

Conclusion: Poor to moderate agreement was found between PSG and each of the tested devices, however Jawbone UP3® had relatively better absolute agreement than other devices in sleep measurements compared with PSG. Consumer grade devices assessed do not have strong enough agreement with gold standard measurement to replace clinical evaluation and PSG sleep testing, however are an opportunity as powerful patient engagement tools for long-term sleep measurement.

Strengths and Limitations of this study

- Consumer grade devices were compared with gold standard in clinic patients.
- More than one device was included for comparison.
- This study includes measure of sleep parameters that clinicians frequently need to review in daily practice, such as total sleep time and sleep efficiency.
- High device failure was found in this study, confirming that consumer grade devices cannot be used to replace high fidelity diagnostic measurement.
- This sample had patients with sleep apnoea, insomnia or hypersomnia as their final sleep diagnosis.

Patient and Public involvement

Patients at our sleep disorders centre sparked the initial interest into assessing the accuracy of consumer-grade sleep monitors. Our clinicians were often asked about the accuracy of home sleep monitors. To answer this question our team invited the patients to be involved in evaluating three commonly available consumer grade devices. Participants were not paid for their involvement but did provide written consent. The findings of this research suggest that consumer-grade sleep monitors can give insights into trends in sleep, but are not accurate enough to replace laboratory measurement.

BACKGROUND

Poor sleep quality and duration has been shown to be an independent risk to overall mortality and for many chronic diseases.[1] The gold standard test for the measurement of sleep and diagnosis of sleep disorders is attended polysomnography (PSG). However, this is an involved and costly test that requires complex equipment, dedicated space, trained staff, and does not lend itself well to multi-night monitoring.

Sales of consumer sleep monitors and wearable consumer-grade smart devices have dramatically increased in recent years, with 33 million units estimated to have been sold in the United States in 2015[2] and the estimated value of the wearable industry in the USA expected to grow to $USD8.5 billion in 2020.[3,4] Consumer-grade devices fall into three major categories (i) wrist based devices (eg Jawbone®, FitBit®); (ii) Bedside devices (eg ResMed S+®, Touch-Free Life Care®); and (iii) Mattress-based devices (eg Beddit®, EarlySense Mattress®, Emfit Bed Sensor®). Each of the categories of devices utilise unique proprietary algorithms for inferring wake/sleep, body position and measures of sleep quality.

The Jawbone UP® (the precursor to the UP3 used in this study) has been compared to PSG in adolescents and concluded to have good agreements for TST, SE and wake after sleep onset (WASO), however the tendency to underestimate TST and sleep efficiency increased with age.[5] In a study of adult women, the FitBitChargeHR® overestimated TST by 27min, and was found to have significantly different SOL and WASO compared to PSG.[5] Similarly in adolescents the Jawbone UP® tended to overestimate TST and SOL, whilst underestimating WASO. The researchers also found greater discrepancies in nights when participants had more disrupted sleep (ie lower TST and greater SOL and WASO).[6]

In patients with suspected central disorders of hypersomnolence, the Jawbone UP3® was found to significantly overestimate TST by an average of 39.6 minutes compared to PSG and was not able to discriminate stages of sleep adequately.[7] Interestingly, the Jawbone UP3® performed similarly to actigraphy in this study. Another clinical study found that the FitBit Flex® overestimated TST more in a group of insomnia patients compared to good sleepers (32.9 mins vs. 6.5 mins).[8] Taken together, these

two studies suggest that consumer-grade sleep devices are less accurate at measuring TST in a clinical sleep disorder population, than they are for good sleepers.

Beddit® in 10 health controls was found to have poor agreement with TST, WASO and SE.[9] SOL was the only measure to have agreement, but had a wide variance.[9] The sensor technology used in the ResMed S+® device has been shown to have moderate accuracy in measuring TST and sleep efficiency in healthy volunteers compared to PSG.[10] Furthermore its utility in measuring sleep disordered breathing has been investigated and found to have reasonable accuracy in detecting moderate obstructive sleep apnoea, with a sensitivity of 89% and specificity 92%.[11]

Patients are increasingly attending sleep clinics with downloads from these devices for discussion with primary care physicians and sleep specialists, or asking clinicians which consumer grade device is best to track sleep at home. These commonly encountered situations in the sleep clinic raise the questions: how reliable are consumer grade devices, and which type of technology is most comparable to gold standard? This study aims to answer these questions with an in-laboratory comparison of PSG with the three consumer devices - Jaw Bone UP3®, Beddit® and ResMed S+® in a sleep clinic population. It was hypothesized that these devices would have similar accuracy in detecting TST, SOL, WASO and SE.

METHODS

Study Population

54 adult patients were consecutively recruited through a private sleep disorders centre in Melbourne, Australia from June 2015 to February 2016. Inclusion criteria were age >18years and any patient who required overnight polysomnography as standard investigation following sleep physician review to either confirm or exclude sleep disordered breathing. All patients attending the laboratory for a polysomnogram were screened for inclusion. Exclusion criteria were age <18years, positive airway pressure titration study, pregnancy and cognitive impairment. Figure 1 demonstrates the consort statement.

Procedure

All assessments took place at an attended sleep laboratory in Melbourne, Australia. The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15). Sleep laboratory staff were trained to set up the 3 devices in addition to regular overnight polysomnography monitoring; lights out time was noted for synchronisation across all devices. The primary outcome measure was Total Sleep Time (TST) and secondary outcomes were sleep onset latency (SOL, min), sleep efficiency (SE, %) as TST/(TST+ total wake time) and wake after sleep onset (WASO, min). Other measures from the consumer grade devices such as time spent in light, deep or rapid eye movement sleep was not compared in this analysis.

Polysomnographic Recording

PSG was measured using a standard six-channel electroencephalography, submental electromyography and electrooculography, electrocardiogram, airflow (thermistor & nasal cannula), respiratory effort, oximetry, snoring (dB sound meter), body position, pulse rate, leg electromyography and digital video, recorded according to American Academy of Sleep Medicine standards.[12] The following standard sleep parameters were recorded via PSG: Total sleep time (TST), sleep onset latency (SOL, min), total wake time (TWT, min), sleep efficiency (SE, %) as TST/(TST+TWT) and wake after sleep onset (WASO, min). Participants were classified as having obstructive sleep apnoea if the apnoea hypopnoea index was >5 events/hr. A single registered polysomnographic technologist scoring the PSG was blinded to the download of consumer grade devices and raw data was scored using Compumedics® amplifiers and Profusion software version 3 (Compumedics®, Abbotsford, Victoria, Australia).

JawBone UP3®

Participants were fitted with the JawBone Up3® on the participant's non dominant wrist with the Jawbone UP3® shortly before lights out time. Data was collected via a dedicated iPod Touch, synced to the Jawbone® app version 4.0.0.[13] This consumer-grade actigraphy device has a three-axis accelerometer and heart rate monitor, which together measure TST, SOL, WASO and SE which were exported by a technician the following morning after the PSG was complete.

ResMed S+®

The ResMed S+® is a non-contact radio-frequency sensor that continuously measures the biomotion due to breathing and body-movement in bed. The sensor operates in a license-free band at 5.8 GHz, emits an average power less than 1 mV and is capable of sensing movement and breathing over a distance ranging from 0.3 to 1.5 meters. The device was positioned by the bedside and synced shortly before lights out time to a dedicated iPod with the ResMed S+® app Version 1.2.1. [14] Measurements from the ResMed S+® were TST, SOL, WASO, SE which were exported by a technician the following morning after the PSG was complete.

Beddit®

The primary sensor in the Beddit® is a piezoelectric 70cm band that was attached to the mattress prior to patients getting into bed. The device detects micro-movements of the chest wall from heartbeats and respiration and uses ballistocardiography to infer sleep stage and time. Ballistocardiography is a non-invasive measurement of cardiac output and respiration by converting mechanical motion (e.g. movement generated by a heartbeat) to a digital signal. Measurements from the Beddit® were taken each night using the device synced to a dedicated iPod running the Beddit® app version 1.[15] Output from the app included TST, SOL, WASO, SE and HR which were exported by a technician the following morning after the PSG was complete.

Statistical analyses

Each of the three non-invasive devices were compared with PSG as the gold standard on an intention to treat basis. The primary and secondary outcomes were compared on total measurements over the night, not epoch-by-epoch method. Summary statistics of the study population are presented. For all normally distributed continuous variables mean and SD, whereas for non-normally distributed variables median and IQR were presented. Normality was assessed using the Shapiro-Wilk's test. Frequencies and proportions are presented for categorical variables. Extent of agreement and reliability between gold standard and each of the selected test devices, was assessed using Intraclass Correlation Coefficients (ICCs) with two-way random-

effects model. Agreement was considered moderate, good and excellent if the ICC values were between 0.5 and 0.75, 0.75 and 0.9 and >0.9 respectively[16].

Additionally, Bland-Altman plots[17] were used to visualize the agreement between gold standard polysomnography and each of the selected devices. The average of two measurements was plotted on x-axis and difference between the two along y-axis. The mean of the differences provided an estimate of average bias between the methods. The upper and lower Limits of agreement (LOA) were calculated which correspond to the mean difference (Gold standard–Selected method) ±2 standard deviations (SD). LOA estimated the interval that a given proportion of differences between the measurements is likely to lie within and will be used to determine if the methods can be used interchangeably. Cohen's d is reported for the magnitude of the effect size. In case of non-normally distributed data, effect size 'r' was calculated by dividing Z statistic by the square root of the sample size (N)). Interpretation of r is 0.10 - < 0.3 (small effect), 0.30 - < 0.5 (moderate effect) and >= 0.5 (large effect)[18]. Data were analysed using R (4.0.4) (https://www.r-project.org/) (R Core Team, 2017).

RESULTS

Fifty-four adult patients (57% females) with a mean age of 48.09 (±SD 18.05) years participated in this study. Table 1 presents demographics of study population. The final sleep diagnosis found was obstructive sleep apnoea in 33 (61%), insomnia 9 (17%) and central hyper-somnolence disorder in 12 (22%) participants. The mean PSG detected TST was 371min (SD±69), SOL of 16min (SD±15), WASO 63min (SD±56) and SE of 82% (SD±13%). The absolute values of the measurements for each device is summarised in table 2.

The results of the Bland-Altman analyses and intra-class correlation are summarized in table 3. For our primary outcome, TST, the mean difference between Jawbone UP3® and PSG was -28.57min indicating that on an average the Jawbone UP3® overestimated TST by 28mins (LOA= -157.37 to 100.23). The magnitude of effect size was small (d=0.4). Generally, PSG measured TST 157mins below or 100mins above Jawbone, however, data were closer between 300-400min. A moderate degree of reliability for recording TST was found between PSG and Jawbone UP3® with

an intra-class correlation coefficient (ICC) of 0.6 and 95% CI from 0.34 to 0.77 (p<0.001). The mean difference in SOL between PSG and Jawbone UP3® was -0.14 min (LOA= -40.23 to 39.95), demonstrating an overestimation of SOL by Jawbone UP3® by 0.14 min. This negative bias seems to be due to measurements over 15 min. The magnitude of difference was small (r=0.2).The reliability between the two methods was between poor to moderate (ICC=0.29; 95% CI=-0.04 to 0.57; p=0.04). Similarly, Jawbone UP3® overestimated WASO by only 1.7min (LOA= -105.71 to 102.32, d=0.03) and bias was seen for measurements over 50 minutes. The agreement between Jawbone UP3® and PSG for WASO was between poor to moderate (ICC=0.55; 95% CI= 0.29-0.73; p <0.001). However, sleep efficiency was underestimated by Jawbone by 0.5% (LOA: -18.96 to 19.99), with bias found with measurements less than 85%. The magnitude of difference was small (d=0.05) The ICC for agreement between Jawbone Up3® and PSG regarding SE was 0.66 (95% CI=0.41 to 0.81; p<0.001) indicating poor to good reliability between the two measures based on 95% CI.

The ResMed S+® underestimated TST by 34 min (LOA= -183.33 to 257.06), with 3 measurements outside the LOA, and magnitude of difference was small (r=0.21). ICC of 0.36 (95% CI: 0.02-0.63; p=0.02) indicating 'poor to moderate' reliability. Conversely, ResMed S+® overestimated SOL by 35.6 min (LOA = -128.89 to 57.68) and effect size was large (r=0.8). The negative bias was seen with measurements greater than 30min. A poor agreement for SOL was seen between the two methods (ICC -0.01 (95% CI:-0.21 – 0.26; p=0.51)). Similarly ResMed S+® recorded WASO 27 min more than PSG (LOA:-127.9 to 73.53 min) and a large effect was found (r=0.52). Reliability between methods was between poor to excellent (ICC= 0.61; 95% CI= 0.28 to 0.8, p <0.01)). Sleep efficiency was underestimated by ResMed S+®, by 15.88 min (LOA=-22.31 to 54.06) and the effect size was large (r=0.8) ICC value of 0.28 (95% CI= -0.06 to 0.58; p=0.06) was found.

The Beddit® and PSG had the least agreement for all outcomes except TST compared to other devices. TST was underestimated by 53min (LOA= -132.01 to 238.79) The magnitude of difference was large (r=0.55) and reliability poor to moderate (ICC= 0.40; 95% CI=0.09 to 0.63; p =0.01). SOL was overestimated by 45min (LOA= -163.33 to 74.09) with a large effect size (r=0.78) and poor reliability (ICC = 0.004 (95%

CI=-0.173 to 0.22; p=0.48). SE was underestimated by 1.35% (LOA= -36.11 to 38.81) with a small effect size (r=0.13) and poor agreement (ICC 0.26;95% CI=-0.04 to 0.51; p=0.06).

Figure 2-4 demonstrates TST agreement for the three devices compared with PSG displayed as Bland-Altman plots and Table 3 summarises the mean differences and intra-class correlation coefficients. Bland-Altman plots for the three devices and SOL, WASO and SE compared with PSG are shown in supplementary figure 1.

Consumer-grade recording failure

Consumer-grade devices were set-up by Sleep Scientist staff each night at the time of the standard PSG set-up. Despite this, device or recording failure resulting in inability to record sufficient data, on the single night of recording, in the consumer-grade devices was common. Failure to synchronise with the dedicated Bluetooth device was the most common reason for device failure. The ResMed S+® failed to synchronise the most, with 25/54 nights (46%) resulting in recording failure. The Jawbone and Beddit had similar rates of synchronisation failure (12/54, 22%), however not usually in the same room or on the same patient. Comparisons were made on an intention to treat analysis, even where large differences in TST were seen.

DISCUSSION

The agreement of these three consumer-grade smart devices have simultaneously been compared with gold standard attended PSG in an adult sleep clinic cohort. For each of the devices, there were components of sleep measurement with poor to moderate agreement with the gold standard. This study found the primary outcome measure of TST was overestimated by, Jawbone UP3® whereas both ResMed S+® and Beddit® underestimated it. The Jawbone UP3® also overestimated SOL and WASO, however the magnitude of difference was very small. Generally Jawbone UP3® had better agreement across all outcomes however for SE agreement was better between ResMed S+® and PSG . The Beddit® had the least

agreement with PSG, all components having poor agreement when compared with gold standard PSG.

Wearable devices, particularly wrist-worn accelerometers have now been widely compared with PSG. Similar to the results of this study, the accelerometers have been shown to overestimate total sleep time by around 20-30minutes, particularly in sleep disordered populations compared with healthy controls.[5,6,8] Previous investigations into consumer grade accelerometers in clinical populations found TST overestimated by 32.9min[8] in a population of 33 insomnia patients and 39min in 43 hyper-somnolence patients[7]. In our study SOL had a large confidence interval, with bias found with measurements over 15min, consistent with findings of a recent systematic review and meta-analysis.[19]

The Beddit® device is one of the least reviewed consumer grade devices out of the three assessed in this study. Tuominen *et al.* (2019) found in 10 healthy controls the Beddit® overestimated total sleep time by 43min, whereas our data suggests a significant underestimation (PSG TST 371min versus Beddit® TST 321 min) with a larger sample size (n = 42). Tuominen *et al.* (2019) was also able to access WASO data, which was not available with the model of Beddit® tested in this study, and found to underestimate WASO by 32min. Non-wearable devices have a potential growing market as non-intrusive home monitors of sleep, as they can be applied in a "set and forget" method. Thus further refinement and evaluation of bed-based devices would be desirable.

The high device synchronisation failure rate in our study is concerning, despite the set-up being performed by sleep laboratory scientific staff. There is no way to calibrate these consumer-grade devices over time and it is difficult to monitor device connectivity to the Bluetooth device until the next morning. The high failure rate further confirms the role of these consumer devices is not to replace that of a diagnostic sleep study.

The main strength of this study was the sample size and that it was conducted in a clinical adult sleep population with a range of suspected sleep disorders. This makes the findings more translatable to clinicians managing patients with sleep disorders. Further, assessing a number of different devices is a novel approach. The weaknesses

of the study include a high device recording failure rate, predominantly with Bluetooth synchronisation failure. Epoch-by-epoch analysis was not performed. Further, sales of devices tested in this study have since been discontinued; however, the technology has been incorporated into subsequent models that are still available.

This study indicates that the wrist worn Jawbone UP3® had the best agreement in measuring sleep compared with gold standard and can provide useful information about commonly measured parameters of sleep quality. For Sleep Medicine Clinicians, the translation of these findings, is that when our patients present with longitudinal measurements of sleep from their consumer grade devices, we can be reassured that wrist worn devices have reasonably accuracy and can be harnessed as an engagement tool for behavioural sleep interventions. This is consistent message with the American Academy of Sleep Medicine's position statement about the use of consumer-grade sleep devices stating that these devices cannot be used for clinical diagnosis, however they allow for meaningful discussions with patients about sleep and encourage active participation in sleep-related health care.[20]

CONCLUSION

Given the large body of literature linking sleep quality to mortality and many chronic diseases, patient-collected longitudinal sleep data provides a powerful insight into a patient's overall health. This study adds to the data of consumer grade wearable sleep monitors, showing they can provide some reliable information compared to gold standard PSG, however do not replace clinical evaluation and gold-standard PSG sleep testing. In reviewing sleep data collected by patients with consumer-grade devices, clinicians are encouraging measurement and quantification of sleep, which in turn will likely emphasise the importance of quality sleep in maintaining good health.

ACKNOWLEDGMENTS

Sleep laboratory staff at St Vincent's Private Hospital, East Melbourne for their set up efforts. Telstra Corporation Ltd (Australia) for the provisions of the Jawbone UP3, ResMed (San Diego) for the ResMed S+ and Beddit Ltd (Finland) for the supply of the

test devices used. The authors acknowledge the statistical support received through the Metro South Health Biostatistics Service.

Ethics Approval

The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15).

Dataset availability

The dataset will be available upon emailed request to the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public or not-for-profit sectors.

Competing Interests

The Telstra Corporation Ltd (Australia) provided the Jawbone UP3 test devices used in the study, ResMed (San Diego) provided the ResMed S+ and Beddit Ltd (Finland) provided the Beddit device. The data reported in this manuscript was presented as a poster at the 27th Annual Scientific Meeting of the Australasian Sleep Association and the Australasian Sleep Technologists Association, held on 22–24 October 2015, in Melbourne, Australia.

Author contributions

Claire M. Ellender	Protocol preparation, Participant consent, Data collection, Data analysis, Manuscript preparation
Syeda Farah Zahir	Data curation, Data analysis, manuscript preparation
Rosemarie Joyce	Participant consent, Data collection, Manuscript preparation
Hailey Meaklim	Data analysis, Manuscript preparation
David Cunnington	Protocol preparation, Data analysis, Manuscript preparation
John Swieca	Protocol preparation, Data analysis, Manuscript preparation

Table 1. Patient demographics

Variable	Results (n = 54)
Age in years, mean (SD)	48.09 (±SD 18.05)
Gender	31 (57%) women
	23 (43%) men
BMI kg/m^2, median (IQR)	27 (24-31)
PSG AHI events/hr, median (IQR)	9 (3-18.75)
Indication for PSG	
Rule in suspected OSA	32 (60%)
Rule out OSA	22 (40%)
Final clinical diagnosis	
OSA syndrome	33 (61%)
Insomnia	9 (17%)
Hypersomnia	12 (22%)

PSG, Polysomnogram; BMI, Body Mass Index; AHI, Apnoea hypopnoea index; OSA, Obstructive sleep apnoea

Table 2 Mean sleep duration

		DEVICE		
VARIABLE	PSG	Jawbone UP3®	ResMed S+®	Beddit®
		(N = 42)	(N = 29)	(N = 42)
TST (MIN SD±)	371 ±69	397 ±83	345.8 ±120	321 ±107
SOL (MIN)	16 ±15	18 ±16	50 ±44	60 ±57
WASO (MIN)	63 ±56	65 ±55	80 ±72	-
SE (%)	82.4 ±13	82.9 ±11	68.8 ±21	81 ±17

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency.

Table 3. Comparison of the outcomes between polysomnography (gold standard) and each of the selected methods

	TST (min)	SOL (min)	WASO (min)	(%)
Jawbone vs PSG Bland-Altman Analysis				
N	42	36	41	35
Bias	-28.57	-0.14	-1.70	0.51
LOA	-157.37 to 100.23	-40.23 to 39.95	-105.71 to 102.32	-18.96 to 19.99
Cohen's d or r (Magnitude)	0.44 (Small)	0.13* (Small)	0.03 (Small)	0.05 (Small)
ICC	0.6 (95% CI= 0.34-0.77; p<0.001)	0.29 (95% CI= -0.04-0.57; p=0.04)	0.55 (95% CI= 0.29-0.73; p <0.001).	0.65 (95% CI=0.41-0.81; p<0.001)
ResMed S+ vs PSG Bland-Altman Analysis				
N	29	29	29	29
Bias	34.36	-35.60	-27.19	15.88
LOA	-188.34 to 257.06	-128.89 to 57.68	-127.91 to 73.53	-22.31 to 54.06
Cohen's d or r (Magnitude)	*0.41 (Moderate)	*0.81 (Large)	*0.52 (Large)	*0.8(Large)
ICC	0.36 (95% CI: 0.02-0.63; p=0.02)	-0.01 (95% CI:- 0.21-0.26; p=0.51)	0.61 (95% CI= 0.28-0.8; p <0.01)	0.06 (95% CI= -0.06-0.58; p=0.06)
Beddit vs PSG. Bland-Altman Analysis				
N	42	42	NA	44
Bias	53.39	-44.62	NA	1.35
LOA	-132.0 to 238.79	-163.33 to 74.09	NA	-36.11 to 38.81

Cohen's d or r (Magnitude)	*0.55(Large)	*0.78(Large)	NA	*0.31 (Small)
ICC	0.40 (95% CI=0.09-0.63; p =0.01)	0.004 (95% CI=-0.173-0.22; p=0.48)	NA	0.26;95% CI=-0.04 to 0.51; p=0.06

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. N=count of pairwise complete cases in groups; LOA=Limits of Agreement (MD±2SD) * effect size=r

Figure captions

Figure 1 CONSORT statement of included participants. CPAP, Continuous Positive Airway pressure.

Figure 2 Bland-Altman plot of the TST recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference (-28.57min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA= -157.37 to 100.23). This shows PSG measured TST 157mins below or 100mins above Jawbone, however, data were closer between 300-400min. TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

Figure 3 Bland-Altman plot of the TST recorded by the ResMed S+® and PSG. The middle line represents the mean difference (34.36min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA= -183.33 to 257.06). TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

Figure 4. Bland-Altman plot of the TST recorded by the Beddit® and PSG. The middle line represents the mean difference (53.39min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA=-132.01 to 238.79). TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

Supplementary Figure 1 Bland-Altman plot of the recorded by PSG compared with the three devices. The middle line represents the mean difference and the upper and lower dotted line represents the upper and lower limits of agreement.

(5a) Sleep onset latency (SOL) Jawbone UP3® compared with PSG

(5b) Sleep onset latency (SOL) ResMed S+® compared with PSG

(5c) Sleep onset latency (SOL) Beddit® compared with PSG

(5d) Wake after sleep onset (WASO) Jawbone UP3® compared with PSG

(5e) Wake after sleep onset (WASO) ResMed S+® compared with PSG

(5f) Sleep efficiency (SE) Jawbone UP3® compared with PSG

(5g) Sleep efficiency (SE) ResMed S+® compared with PSG

(5h) Sleep efficiency (SE) Beddit® compared with PSG

References

1. Cai H, Shu XO, Xiang YB, Yang G, Li H, Ji BT, et al. Sleep duration and mortality: a prospective study of 113 138 middle-aged and elderly Chinese men and women. Sleep 2015;38(4):529-36.
2. Davona T. The Wearables Report: Growth Trends, Consumer Attitudes and Why Smart Watches Will Dominate, Business Insider, February 12, 2015. Buisness Insider Australia. 2015; JUL 6 2015. Available from: http://www.businessinsider.com.au/the-wearable-computing-market-report-bii-2015-7
3. de Zambotti M, Baker FC, Willoughby AR, Godino JG, Wing D, Patrick K, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. Physiology & behavior 2016;158:143-9.
4. Intelligence C. US Enterprise Wearables Market: 5-Year Forecast, 2014–2019. 2015. Available from: http://www.marketwired.com/press-release/compass-intelligence-forecasts-wearables-enterprise-grow-exponentially-us-device-revenue-2032309.htm
5. de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults. Chronobiol Int 2015;32(7):1024-8.
6. de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents. Sleep 2015;38(9):1461-8.
7. Cook JD, Prairie ML, Plante DT. Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy. J Clin Sleep Med 2018;14(5):841-8.
8. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. J Psychosom Res 2017;97:38-44.

9. Tuominen J, Peltola K, Saaresranta T, Valli K. Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. J Clin Sleep Med 2019;15(3):483-7.
10. De Chazal P, Fox N, O'Hare E, Heneghan C, Zaffaroni A, Boyle P, et al. Sleep/wake measurement using a non-contact biomotion sensor. J Sleep Res 2011;20(2):356-66.
11. Zaffaroni A, de Chazal P, Heneghan C, Boyle P, Mppm PR, McNicholas WT. SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index. Conf Proc IEEE Eng Med Biol Soc 2009;2009:7091-4.
12. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. J Clin Sleep Med 2012;8(5):597-619.
13. AliphCom dba Jawbone. Jawbone UP3. San Francisco2016 [cited 2016 19/4/16]. Available from: https://jawbone.com/support/articles/000001027/download-the-app
14. ResMed. San Diego2016 [cited 2016 19/4/16]. Available from: https://itunes.apple.com/us/app/s+-by-resmed/id883611019?mt=8
15. Beddit Ltd. Beddit Sleep Tracker. Helsinki, Finnland2016 [cited 2016 19/4/16]. Available from: http://support.beddit.com/hc/en-us/articles/201422237-Downloading-the-Beddit-app
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine 2016;15(2):155-63.
17. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet 1986;1(8476):307-10.
18. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends in Sport Sciences 2014;21:19-25.
19. Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. Sleep Med Rev 2020;49:101227.

20. Khosla S, Deak MC, Gault D, Goldstein CA, Hwang D, Kwon Y, et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. J Clin Sleep Med 2018;14(5):877-80.

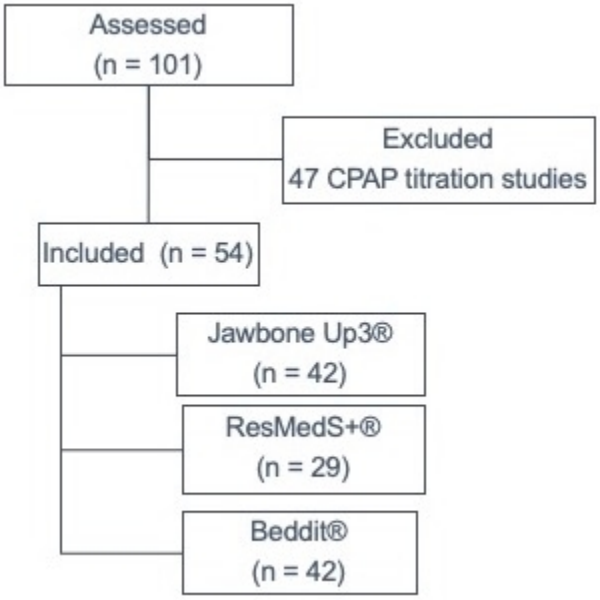


Figure 1 - Consort statement of data collection.

156x153mm (54 x 54 DPI)

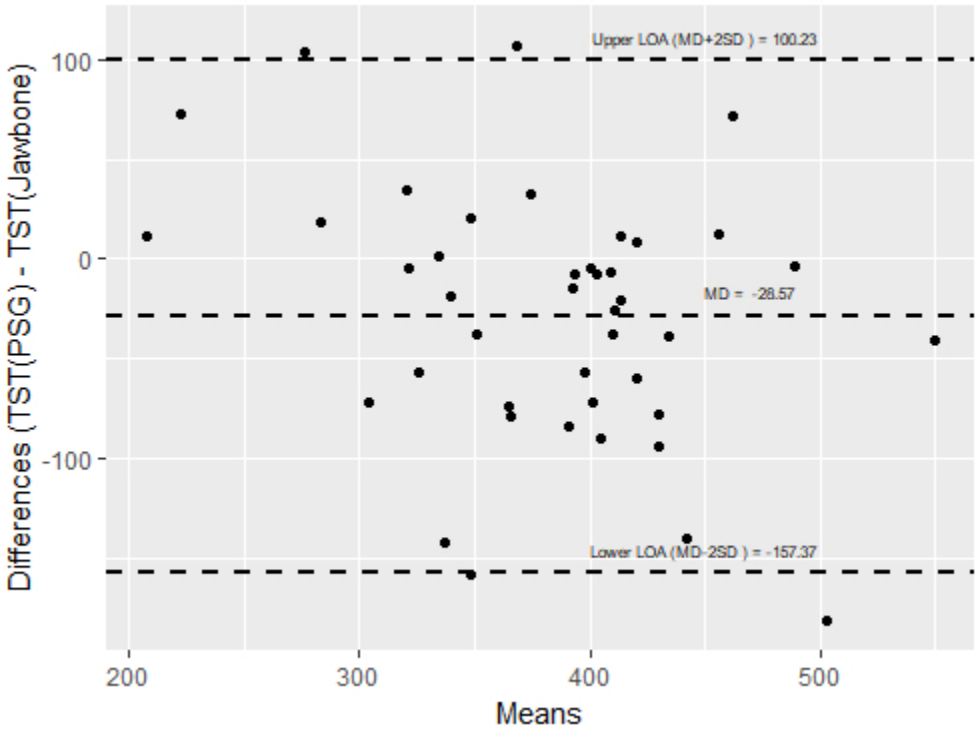


Figure 2 Bland-Altman plot of the TST recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference (-28.57min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA= -157.37 to 100.23). This shows PSG measured TST 157mins below or 100mins above Jawbone, however, data were closer between 300-400min. TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

127x101mm (100 x 100 DPI)

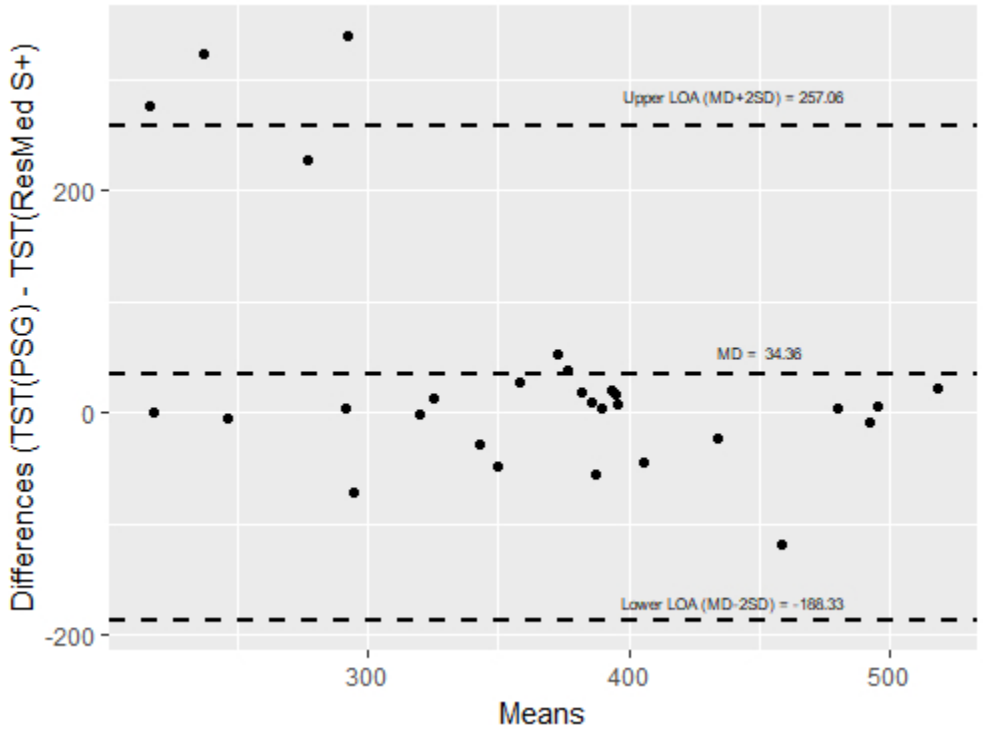


Figure 3 Bland-Altman plot of the TST recorded by the ResMed S+® and PSG. The middle line represents the mean difference (34.36min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA= -183.33 to 257.06). TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

127x101mm (100 x 100 DPI)

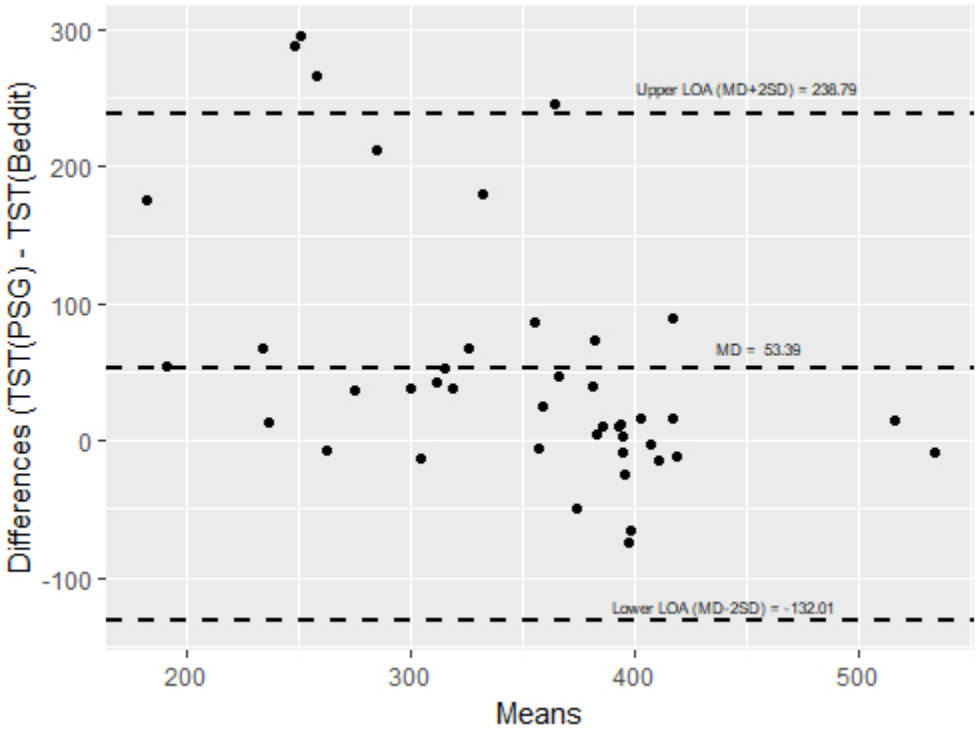
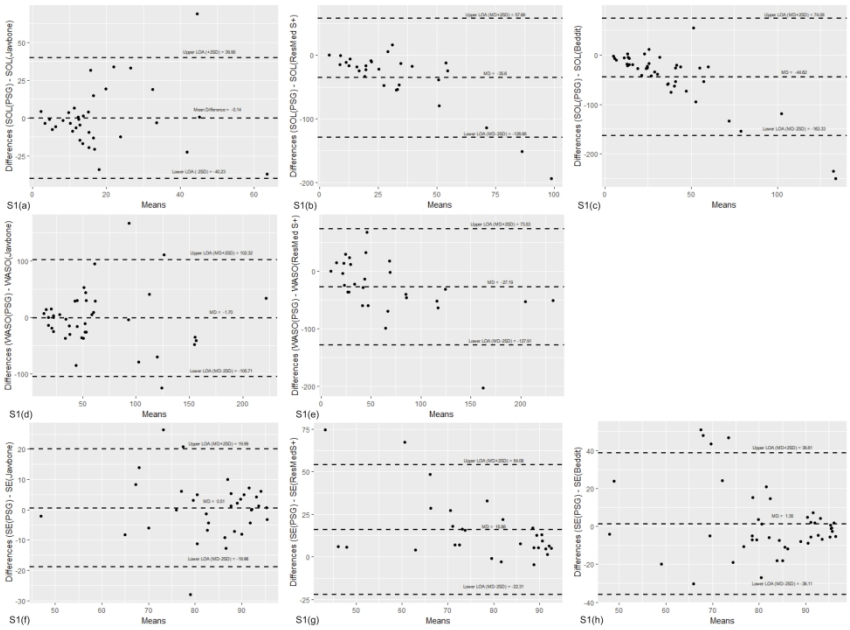


Figure 4. Bland-Altman plot of the TST recorded by the Beddit® and PSG. The middle line represents the mean difference (53.39min) and the upper and lower dotted line represents the upper and lower limits of agreement (LOA=-132.01 to 238.79). TST, total sleep time; PSG, polysomnography; LOA, limits of agreement.

127x101mm (100 x 100 DPI)



148x104mm (600 x 600 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	1, 5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	
Participants	**6**	Eligibility criteria	5-7
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	5-7
	8	Where and when potentially eligible participants were identified (setting, location and dates)	5-7
	9	Whether participants formed a consecutive, random or convenience series	5
Test methods	**10a**	Index test, in sufficient detail to allow replication	5
	10b	Reference standard, in sufficient detail to allow replication	6
	11	Rationale for choosing the reference standard (if alternatives exist)	7
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-8
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	6
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	6
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	7-8
	15	How indeterminate index test or reference standard results were handled	7
	16	How missing data on the index test and reference standard were handled	10
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	5
	18	Intended sample size and how it was determined	-
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Figure1
	20	Baseline demographic and clinical characteristics of participants	Table 1
	21a	Distribution of severity of disease in those with the target condition	Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	-
	22	Time interval and any clinical interventions between index test and reference standard	-
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Table 2
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Table 3
	25	Any adverse events from performing the index test or the reference standard	-
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-12
	27	Implications for practice, including the intended use and clinical role of the index test	12
OTHER INFORMATION			
	28	Registration number and name of registry	-
	29	Where the full study protocol can be accessed	-
	30	Sources of funding and other support; role of funders	13

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-044015.R2
Article Type:	Original research
Date Submitted by the Author:	04-Aug-2021
Complete List of Authors:	Ellender, Claire; Princess Alexandra Hospital, ; The University of Queensland, Zahir, Syeda ; The University of Queensland, QCIF Facility for Advanced Bioinformatics Meaklim, Hailey; Melbourne Sleep Disorders Centre; St Vincent's Hospital Melbourne Pty Ltd Joyce, Rosemarie; Melbourne Sleep Disorders Centre Cunnington, David; Melbourne Sleep Disorders Centre, Swieca, John; Melbourne Sleep Disorders Centre
Primary Subject Heading:	Respiratory medicine
Secondary Subject Heading:	General practice / Family practice
Keywords:	SLEEP MEDICINE, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, RESPIRATORY MEDICINE (see Thoracic Medicine), Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ORIGINAL RESEARCH

Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population

Claire M. Ellender[1,2]
Syeda Farah Zahir[3]
Hailey Meaklim[1]
Rosemarie Joyce[1]
David Cunnington[1]
John Swieca[1]
[1] Melbourne Sleep Disorders Centre, East Melbourne, Victoria, Australia
[2] Princess Alexandra Hospital, Brisbane, Australia
[3] QCIF Facility for Advanced Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, 4072, AUSTRALIA
ORCID's: Hailey Meaklim 0000-0003-0448-3567; Claire Ellender 0000-0002-1727-576X; David Cunnington 0000-0002-8403-0420; Syeda Farah Zahir 0000-0002-2074-6999;
John Swieca 0000-0001-8281-4048.

Correspondence: Dr Claire Ellender
Department of Respiratory & Sleep Medicine, Princess Alexandra Hospital, 199 Ipswich Rd, Brisbane, Qld, Australia 4170
Tel +61731762698, Fax +61 731766170
Email Claire.ellender@health.qld.gov.au

Word Count: 3180

Keywords: Consumer sleep monitor; accelerometer, polysomnography; sleep; sleep measurement; validation; wearables

ABSTRACT:

Objectives: Consumer-grade smart devices are now commonly used by the public to measure waking activity and sleep. However, the ability of these devices to accurately measure sleep in clinical populations warrants more examination. The aim of the present study was to assess the accuracy of three consumer-grade sleep monitors compared with gold standard polysomnography (PSG).

Design: A prospective cohort study was performed.

Setting: Adults undergoing PSG for investigation of a suspected sleep disorder.

Participants: 54 sleep-clinic patients were assessed using three consumer-grade sleep monitors (Jawbone UP3®, ResMed S+® and Beddit®) in addition to PSG.

Outcomes: Jawbone UP3®, ResMed S+® and Beddit® were compared with gold standard in-laboratory polysomnography on 4 major sleep parameters - total sleep time (TST), sleep onset latency (SOL), Wake After Sleep Onset (WASO) and sleep efficiency (SE).

Results: The accelerometer Jawbone UP3® was found to overestimate TST by 28mins (limits of agreement, LOA= -100.23 to 157.37), with reasonable agreement compared with gold standard for TST, WASO and SE. The doppler radar ResMed S+® device underestimated TST by 34mins (LOA= -257.06 to 188.34) and had poor absolute agreement compared with PSG for TST, SOL and SE. The mattress device, Beddit® underestimated TST by 53mins (LOA= -238.79 to 132) on average and poor reliability compared to PSG for all measures except TST. High device synchronisation failure occurred, with 20% of recordings incomplete due to Bluetooth drop out and recording loss.

Conclusion: Poor to moderate agreement was found between PSG and each of the tested devices, however Jawbone UP3® had relatively better absolute agreement than other devices in sleep measurements compared with PSG. Consumer grade devices assessed do not have strong enough agreement with gold standard measurement to replace clinical evaluation and PSG sleep testing. The models tested here have been superseded and newer models may have increase accuracy and thus potentially powerful patient engagement tools for long-term sleep measurement.

Strengths and Limitations of this study

- Consumer grade devices were compared with gold standard in clinic patients.
- More than one device was included for comparison.
- This study includes measure of sleep parameters that clinicians frequently need to review in daily practice, such as total sleep time and sleep efficiency.
- High device failure was found in this study, confirming that consumer grade devices cannot be used to replace high fidelity diagnostic measurement.
- This sample had patients with sleep apnoea, insomnia or hypersomnia as their final sleep diagnosis.

BACKGROUND

Poor sleep quality and duration has been shown to be an independent risk to overall mortality and for many chronic diseases.[1] The gold standard test for the measurement of sleep and diagnosis of sleep disorders is attended polysomnography (PSG). However, this is an involved and costly test, that requires complex equipment, dedicated space, trained staff, and does not lend itself well to multi-night monitoring.

Sales of consumer sleep monitors and wearable consumer-grade smart devices have dramatically increased in recent years, with 33 million units estimated to have been sold in the United States in 2015[2] and the estimated value of the wearable industry in the USA expected to grow to $USD8.5 billion in 2020.[3 4] Consumer-grade devices fall into three major categories (i) wrist based devices (eg Jawbone®, FitBit®); (ii) Bedside devices (eg ResMed S+®, Touch-Free Life Care®); and (iii) Mattress-based devices (eg Beddit®, EarlySense Mattress®, Emfit Bed Sensor®). Each of the categories of devices utilise unique proprietary algorithms for inferring wake/sleep, body position and measures of sleep quality.

The Jawbone UP® (the precursor to the UP3 used in this study) has been compared to PSG in adolescents and concluded to have good agreements for TST, SE and wake after sleep onset (WASO), however the tendency to underestimate TST and sleep efficiency increased with age.[5] In a study of adult women, the FitBitChargeHR® overestimated TST by 27min, and was found to have significantly different SOL and WASO compared to PSG.[5] Similarly in adolescents the Jawbone UP® tended to overestimate TST and SOL, whilst underestimating WASO. The researchers also found greater discrepancies in nights when participants had more disrupted sleep (i.e., lower TST and greater SOL and WASO).[5] In patients with suspected central disorders of hypersomnolence, the Jawbone UP3® was found to significantly overestimate TST by an average of 39.6 minutes compared to PSG and was not able to discriminate stages of sleep adequately.[6] Interestingly, the Jawbone UP3® performed similarly to actigraphy in this study. Another clinical study found that the FitBit Flex® overestimated TST more in a group of insomnia patients compared to good sleepers (32.9 mins vs. 6.5 mins).[7] Taken together, these two studies suggest that consumer-grade sleep devices are less accurate at measuring TST in a clinical sleep disorder population, than they are for good sleepers.

The Beddit® mattresses based device has been found in 10 health controls to have poor agreement with TST (overestimated by 43.5min), WASO and SE.[8] SOL was the only measure to have agreement, but had a wide variance.[8] The sensor technology used in the ResMed S+® device has been shown to have moderate accuracy in measuring TST and sleep efficiency in

healthy volunteers compared to PSG.[9] Furthermore its utility in measuring sleep disordered breathing has been investigated and found to have reasonable accuracy in detecting moderate obstructive sleep apnoea, with a sensitivity of 89% and specificity 92%.[10]

Patients are increasingly attending sleep clinics with downloads from consumer-grade devices for discussion with primary care physicians and sleep specialists. These commonly encountered situations in the sleep clinic raise the questions: how reliable are consumer-grade devices, and which type of technology is most comparable to gold standard? This study aims to answer these questions with an in-laboratory comparison of PSG with the three consumer devices - Jaw Bone UP3®, Beddit® and ResMed S+® in a sleep clinic population. It was hypothesized that these devices would have similar accuracy in detecting TST, SOL, WASO and SE.

METHODS

Study Population

54 adult patients were consecutively recruited through a private sleep disorders centre in Melbourne, Australia from June 2015 to February 2016. Inclusion criteria were age >18years and any patient who required overnight polysomnography as standard investigation following sleep physician review to either confirm or exclude sleep disordered breathing. All patients attending the laboratory for a polysomnogram were screened for inclusion. Exclusion criteria were age <18years, positive airway pressure titration study, pregnancy and cognitive impairment. Figure 1 demonstrates the consort statement.

Procedure

All assessments took place at an attended sleep laboratory in Melbourne, Australia. The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15). Sleep laboratory staff were trained to set up the 3 devices in addition to regular overnight polysomnography monitoring; lights out time was noted for synchronisation across all devices. The primary outcome measure was Total Sleep Time (TST) and secondary outcomes were sleep onset latency (SOL, min), sleep efficiency (SE, %) as TST/(TST+ total wake time) and wake after sleep onset (WASO, min). Other measures from the consumer grade devices such as time spent in light, deep or rapid eye movement sleep was not compared in this analysis.

Polysomnographic Recording

PSG was measured using a standard six-channel electroencephalography, submental electromyography and electrooculography, electrocardiogram, airflow (thermistor & nasal cannula), respiratory effort, oximetry, snoring (dB sound meter), body position, pulse rate, leg electromyography and digital video, recorded according to American Academy of Sleep Medicine standards.[11] The following standard sleep parameters were recorded via PSG: Total sleep time (TST), sleep onset latency (SOL, min), total wake time (TWT, min), sleep efficiency (SE, %) as TST/(TST+TWT) and wake after sleep onset (WASO, min). Participants were classified as having obstructive sleep apnoea if the apnoea hypopnoea index was >5 events/hr. A single registered polysomnographic technologist scoring the PSG was blinded to the download of consumer grade devices and raw data was scored using Compumedics® amplifiers and Profusion software version 3 (Compumedics®, Abbotsford, Victoria, Australia).

JawBone UP3®

Participants were fitted with the JawBone Up3® on the participant's non dominant wrist with the Jawbone UP3® shortly before lights out time. Data was collected via a dedicated iPod Touch, synced to the Jawbone® app version 4.0.0.[12] This consumer-grade actigraphy device has a three-axis accelerometer and heart rate monitor, which together measure TST, SOL, WASO and SE which were exported by a technician the following morning after the PSG was complete.

ResMed S+®

The ResMed S+® is a non-contact radio-frequency sensor that continuously measures the biomotion due to breathing and body-movement in bed. The sensor operates in a license-free band at 5.8 GHz, emits an average power less than 1 mV and is capable of sensing movement and breathing over a distance ranging from 0.3 to 1.5 meters. The device was positioned by the bedside and synced shortly before lights out time to a dedicated iPod with the ResMed S+® app Version 1.2.1. [13] Measurements from the ResMed S+® were TST, SOL, WASO, SE which were exported by a technician the following morning after the PSG was complete.

Beddit®

The primary sensor in the Beddit® is a piezoelectric 70cm band that was attached to the mattress prior to patients getting into bed. The device detects micro-movements of the chest wall from heartbeats and respiration and uses ballistocardiography to infer sleep stage and time. Ballistocardiography is a non-invasive measurement of cardiac output and respiration by

converting mechanical motion (e.g. movement generated by a heartbeat) to a digital signal. Measurements from the Beddit® were taken each night using the device synced to a dedicated iPod running the Beddit® app version 1.[14] Output from the app included TST, SOL, WASO, SE and HR which were exported by a technician the following morning after the PSG was complete.

Statistical analyses

Each of the three non-invasive devices were compared with PSG as the gold standard on an intention to treat basis. The primary and secondary outcomes were compared on total measurements over the night, not epoch-by-epoch method. Summary statistics of the study population are presented. For all normally distributed continuous variables mean and SD, whereas for non-normally distributed variables median and IQR were presented. Normality was assessed using the Shapiro-Wilk's test. Frequencies and proportions are presented for categorical variables. Extent of agreement and reliability between gold standard and each of the selected test devices, was assessed using Intraclass Correlation Coefficients (ICCs) with two-way random-effects model. Agreement was considered moderate, good and excellent if the ICC values were between 0.5 and 0.75, 0.75 and 0.9 and >0.9 respectively[15].

Additionally, Bland-Altman plots[16] were used to visualize the agreement between gold standard polysomnography and each of the selected devices. The average of two measurements was plotted on x-axis and difference between the two along y-axis. The mean of the differences provided an estimate of average bias between the methods. The upper and lower Limits of agreement (LOA) were calculated which correspond to the mean difference (Gold standard–Selected method) ±2 standard deviations (SD). LOA estimated the interval that a given proportion of differences between the measurements is likely to lie within and will be used to determine if the methods can be used interchangeably. Cohen's d is reported for the magnitude of the effect size. In case of non-normally distributed data, effect size 'r' was calculated by dividing Z statistic by the square root of the sample size (N)). Interpretation of r is 0.10 - < 0.3 (small effect), 0.30 - < 0.5 (moderate effect) and >= 0.5 (large effect)[17]. Data were analysed using R (4.0.4) (https://www.r-project.org/) (R Core Team, 2017).

Patient and Public involvement

Patients at our sleep disorders centre sparked the interest to assess the accuracy of consumer-grade sleep monitors. Our clinicians were often asked about the accuracy of home sleep monitors. To answer this question our team invited the patients to be involved in

evaluating three commonly available consumer-grade smart devices. Participants were not paid for their involvement but did provide written consent. The findings of this research suggest that consumer-grade sleep monitors can give insights into trends in sleep but are not accurate enough to replace laboratory measurement.

RESULTS

Fifty-four adult patients (57% females) with a mean age of 48.09 (±SD 18.05) years participated in this study. Table 1 presents demographics of study population. The final sleep diagnosis found was obstructive sleep apnoea in 33 (61%), insomnia 9 (17%) and central hypersomnolence disorder in 12 (22%) participants. The mean PSG detected TST was 371min (SD±69), SOL of 16min (SD±15), WASO 63min (SD±56) and SE of 82% (SD±13%). The absolute values of the measurements for each device are summarised in table 2. The results of the Bland-Altman analyses and intra-class correlation are summarized in table 3 and displayed in figures 2-4.

JawBone UP3®

On average JawBone UP3® overestimated TST by 28.57mins (LOA= -100.23 to 157.37). By inspecting the Bland-Altman plots (shown in Figure 2A), the cluster of points surrounded the mean tightly between 300-400 minutes and there was greater variability with TST below 300min and above 400 minutes. The magnitude of effect size was small (d=0.44). A moderate degree of reliability for recording TST was found between PSG and Jawbone UP3® with an intra-class correlation coefficient (ICC) of 0.6 (95% CI = 0.34 to 0.77; p<0.001).

Bland Altman plot (Figure 2B) suggests that the mean difference in SOL between two methods was very small and on average JawBone UP3® measured SOL 0.14mins (LOA= -39.95 to 40.23) more than the gold standard. The cluster of points surrounded the mean tightly on the left, with greater variability for values over 20 minutes. The magnitude of difference was small (r=0.13). The reliability between the two methods was between poor to moderate (ICC=0.29; 95% CI=-0.04 to 0.57; p=0.04).

Jawbone UP3® overestimated WASO only slightly, 1.7min (LOA= -102.32 to 105.71, d=0.03) compared with PSG. Greater variability was seen for measurements over 50 minutes (as shown in Figure 2C), indicating better estimation of WASO by JawBone UP3® at lower values. The agreement between Jawbone UP3® and PSG for WASO was poor to moderate (ICC=0.55; 95% CI= 0.29-0.73; p <0.001).

The mean difference in SE between two methods indicated that on an average JawBone UP3® measures SE 0.51% (LOA: -18.96 to 19.99) less than the gold standard. This bias seems to be due to measurements less than 85%, with better estimation of SE by JawBone UP3® at higher SE, as seen in Figure 2D. The magnitude of difference was small (d=0.05) The ICC for agreement between Jawbone Up3® and PSG regarding SE was 0.66 (95% CI=0.41 to 0.81; p<0.001) indicating poor to good reliability between the two measures based on 95% CI.

ResMed S+®

As shown in Figure 3A, on average ResMed S+® underestimated TST by 34 minutes (CI: -257 min to 188 min). The mean difference between ResMed S+® measured and PSG measured TST was offset (lying below) zero, suggesting a bias. The points remained in the same general pattern for all x-axis values, except for few outliers at lower mean values. The magnitude of difference was moderate (r=0.4). ICC of 0.36 (95% CI: 0.02-0.63; p=0.02) indicating poor to moderate reliability.

Conversely, ResMed S+® overestimated SOL by 35.6 min (LOA = -57.68 to -128.89) and effect size was large (r=0.8). Cluster of points go from below the mean at short SOL, to above the mean with increasing SOL, showing proportional error, suggesting overestimation of SOL by ResMed S+® at increasing SOL duration, as shown in Figure 3B. A poor agreement for SOL was seen between the two methods (ICC= -0.01; 95% CI: -0.21 – 0.26; p=0.51).

Similarly, ResMed S+® recorded WASO 27 min more than PSG (LOA= -73.53 to 127.91) and a large effect was found (r=0.52). Visual inspection of Bland Altman plot (Figure 3C) suggested that ResMed S+® increasingly overestimating WASO with increasing time. Reliability between methods was between poor to excellent (ICC= 0.61; 95% CI= 0.28 to 0.8, p <0.01).

Visual inspection of the Bland-Altman plot Figure 3D suggests that on average ResMed S+® underestimated SE by 16% (LOA=-54.06 to 22.31). The effect size was large (r=0.8) and an ICC value of 0.28 (95% CI= -0.06 to 0.58; p=0.06) was found. Moreover, the mean difference was not constant, with greater variability at lower values (particularly below 80%), showing proportional bias.

Beddit®

The Beddit® and PSG had the least agreement for all outcomes except TST compared to other devices. TST was underestimated by 53min (LOA= -238.79 to 132). As demonstrated in Figure 4A, the cluster of points shifted from below mean to above mean with increasing TST,

showing a proportional error depending on the duration of sleep. The magnitude of difference was large (r=0.55) and reliability poor to moderate (ICC= 0.40; 95% CI=0.09 to 0.63; p =0.01).

SOL was overestimated by 45min (LOA= -74.09 to 163.33) by the Beddit® compared with PSG. The points were tightly clustering above the mean, and go from above, to below the mean, from left to right (Figure 4B), showing error proportional to the duration of SOL. The effect size was large (r=0.78) and reliability poor (ICC = 0.004; 95% CI=-0.173 to 0.22; p=0.48).

Beddit® slightly underestimated SE by 1.35% (LOA= -38.81 to 36.11). As shown in Figure 4C, variability of points was constant around the mean at values below 80%. This suggest that at higher values, Beddit® estimated SE more closely to the PSG gold standard. The effect size was small (r=0.13) and poor agreement (ICC 0.26; 95% CI=-0.04 to 0.51; p=0.06).

Consumer-grade recording failure

Consumer-grade devices were set-up by Sleep Scientist staff each night at the time of the standard PSG set-up. Despite this, device or recording failure resulting in inability to record sufficient data, on the single night of recording, in the consumer-grade devices was common. Failure to synchronise with the dedicated Bluetooth device was the most common reason for device failure. The ResMed S+® failed to synchronise the most, with 25/54 nights (46%) resulting in recording failure. The Jawbone and Beddit® had similar rates of synchronisation failure (12/54, 22%), however not usually in the same room or on the same patient. Comparisons were made on an intention to treat analysis, even where large differences in TST were seen.

DISCUSSION

The agreement of these three consumer-grade smart devices have simultaneously been compared with gold standard attended PSG in an adult sleep clinic cohort. For each of the devices, there were components of sleep measurement with poor to moderate agreement with the gold standard. This study found the primary outcome measure of TST was overestimated by, Jawbone UP3® whereas both ResMed S+® and Beddit® underestimated it. The Jawbone UP3® also overestimated SOL and WASO, however the magnitude of difference was very small. Generally Jawbone UP3® had better agreement across all outcomes, however for SE

agreement was better between ResMed S+® and PSG . The Beddit® had the least agreement with PSG, all components having poor agreement when compared with gold standard PSG.

Wearable devices, particularly wrist-worn accelerometers have now been widely compared with PSG. Similar to the results of this study, the accelerometers have been shown to overestimate total sleep time by around 20-30minutes, particularly in sleep disordered populations compared with healthy controls.[5 7 18] Previous investigations into consumer grade accelerometers in clinical populations found TST overestimated by 32.9min[7] in a population of 33 insomnia patients and 39min in 43 hyper-somnolence patients[6]. In our study SOL had a large confidence interval, with bias found with measurements over 15min, consistent with findings of a recent systematic review and meta-analysis.[19]

The Beddit® device and mattress devices in general are one of the least studied consumer grade devices. Tuominen *et al.* (2019) found in 10 healthy controls the Beddit® overestimated total sleep time by 43min, whereas our data suggests a significant underestimation (PSG TST 371min versus Beddit® TST 321 min) with a larger sample size (n = 42). Tuominen *et al.* (2019) was also able to access WASO data, which was not available with the model of Beddit® tested in this study and found to underestimate WASO by 32min. Non-wearable devices have a potential growing market as non-intrusive home monitors of sleep, as they can be applied in a “set and forget” method. Thus, further refinement and evaluation of bed-based devices would be desirable.

The high device synchronisation failure rate in our study is concerning, despite the set-up being performed by sleep laboratory scientific staff. There is no way to calibrate these consumer-grade devices over time and it is difficult to monitor device connectivity to the Bluetooth device until the next morning. The high failure rate further confirms the role of these consumer devices is not to replace that of a diagnostic sleep study.

The main strength of this study was the sample size and that it was conducted in a clinical adult sleep population with a range of suspected sleep disorders. This makes the findings more translatable to clinicians managing patients with sleep disorders. Further, assessing a number of different devices is a novel approach. The weaknesses of the study include a high device recording failure rate, predominantly with Bluetooth synchronisation failure. Epoch-by-epoch analysis was not performed. Further, sales of devices tested in this study have since been discontinued. Beddit® was acquired by Apple Inc in May 2017 and relaunched an updated device, the Beddit® 3.5 which has reportedly improved integration with mobile phone health kits[20]. The ResMed S+® was discontinued and subsequently a similar device was launched in 2017 as SleepScore labs, which is similarly Apple iOSS and Adroid

integrated[21]. JawBone® however has gone into liquidation with no subsequent models leading on from the UP3® device[22].

This study indicates that the wrist worn Jawbone UP3® had the best agreement in measuring sleep compared with gold standard and can provide useful information about commonly measured parameters of sleep quality. For Sleep Medicine Clinicians, the translation of these findings, is that when our patients present with longitudinal measurements of sleep from their consumer grade devices, we can be reassured that wrist worn devices have reasonably accuracy and can be harnessed as an engagement tool for behavioural sleep interventions. This is consistent message with the American Academy of Sleep Medicine's position statement about the use of consumer-grade sleep devices stating that these devices cannot be used for clinical diagnosis, however they allow for meaningful discussions with patients about sleep and encourage active participation in sleep-related health care.[23]

CONCLUSION

Given the large body of literature linking sleep quality to mortality and many chronic diseases, patient-collected longitudinal sleep data provides a powerful insight into a patient's overall health. This study adds to the data of consumer grade wearable sleep monitors, showing they can provide some reliable information compared to gold standard PSG, however do not replace clinical evaluation and gold-standard PSG sleep testing. In reviewing sleep data collected by patients with consumer-grade devices, clinicians are encouraging measurement and quantification of sleep, which in turn will likely emphasise the importance of quality sleep in maintaining good health.

ACKNOWLEDGMENTS

Sleep laboratory staff at St Vincent's Private Hospital, East Melbourne for their set up efforts. Telstra Corporation Ltd (Australia) for the provisions of the Jawbone UP3, ResMed (San Diego) for the ResMed S+ and Beddit Ltd (Finland) for the supply of the test devices used. The authors acknowledge the statistical support received through the Metro South Health Biostatistics Service.

Ethics Approval

The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15).

Dataset availability

The dataset will be available upon emailed request to the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public or not-for-profit sectors.

Competing Interests

The Telstra Corporation Ltd (Australia) provided the Jawbone UP3 test devices used in the study, ResMed (San Diego) provided the ResMed S+ and Beddit Ltd (Finland) provided the Beddit device.

Author contributions

Claire M. Ellender	Protocol preparation, Participant consent, Data collection, Data analysis, Manuscript preparation
Syeda Farah Zahir	Data curation, Data analysis, manuscript preparation
Rosemarie Joyce	Participant consent, Data collection, Manuscript preparation
Hailey Meaklim	Data analysis, Manuscript preparation
David Cunnington	Protocol preparation, Data analysis, Manuscript preparation
John Swieca	Protocol preparation, Data analysis, Manuscript preparation

Table 1. Patient demographics

Variable	Results (n = 54)
Age in years, mean (SD)	48.09 (±SD 18.05)
Gender	31 (57%) women
	23 (43%) men
BMI kg/m^2, median (IQR)	27 (24-31)
PSG AHI events/hr, median (IQR)	9 (3-18.75)
Indication for PSG	
Rule in suspected OSA	32 (60%)
Rule out OSA	22 (40%)
Final clinical diagnosis	
OSA syndrome	33 (61%)
Insomnia	9 (17%)
Hypersomnia	12 (22%)

PSG, Polysomnogram; BMI, Body Mass Index; AHI, Apnoea hypopnoea index; OSA, Obstructive sleep apnoea

Table 2 Mean sleep duration

		DEVICE		
VARIABLE	PSG	Jawbone UP3® (N = 42)	ResMed S+® (N = 29)	Beddit® (N = 42)
TST (MIN SD±)	371 ±69	397 ±83	345.8 ±120	321 ±107
SOL (MIN)	16 ±15	18 ±16	50 ±44	60 ±57
WASO (MIN)	63 ±56	65 ±55	80 ±72	-
SE (%)	82.4 ±13	82.9 ±11	68.8 ±21	81 ±17

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency.

Table 3. Comparison of the outcomes between polysomnography (gold standard) and each of the selected methods

	TST (min)	SOL (min)	WASO (min)	(%)
Jawbone vs PSG Bland-Altman Analysis				
N	42	36	41	35
Bias	28.57	0.14	1.70	-0.51
LOA	-100.23 to 157.37	-39.95 to 40.23	-102.32 to 105.71	- 19.99 to 18.96
Cohen's d or r (Magnitude)	0.44 (Small)	0.13* (Small)	0.03 (Small)	0.05 (Small)
ICC	0.6 (95% CI= 0.34-0.77; p<0.001)	0.29 (95% CI= -0.04-0.57; p=0.04)	0.55 (95% CI= 0.29-0.73; p <0.001).	0.65 (95% CI=0.41-0.81; p<0.001)
ResMed S+ vs PSG Bland-Altman Analysis				
N	29	29	29	29
Bias	-34.36	35.60	27.19	-15.88
LOA	-257.06 to 188.34	-57.68 to -128.89	-73.53 to 127.91	-54.06 to 22.31
Cohen's d or r (Magnitude)	*0.41 (Moderate)	*0.81 (Large)	*0.52 (Large)	*0.8(Large)
ICC	0.36 (95% CI: 0.02-0.63; p=0.02)	-0.01 (95% CI:-0.21-0.26; p=0.51)	0.61 (95% CI= 0.28-0.8; p <0.01)	0.06 (95% CI= -0.06-0.58; p=0.06)
Beddit vs PSG. Bland-Altman Analysis				
N	42	42	NA	44
Bias	-53.39	44.62	NA	-1.35
LOA	-238.79 to 132	-74.09 to 163.33	NA	-38.81 to 36.11
Cohen's d or r (Magnitude)	*0.55(Large)	*0.78(Large)	NA	*0.31 (Small)
ICC	0.40 (95% CI=0.09-0.63; p	0.004 (95% CI=-0.173-0.22;	NA	0.26;95% CI=-0.04 to 0.51;

	=0.01)	p=0.48)	p=0.06

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. N=count of pairwise complete cases in groups; LOA=Limits of Agreement (MD±2SD) * effect size=r. Bias = the mean differences between test device minus PSG.

Figure captions

Figure 1 CONSORT statement of included participants. CPAP, Continuous Positive Airway pressure.

Figure 2 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

Figure 3 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the ResMed S+® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

Figure 4. Bland-Altman plot of three outcomes (TST, SOL and SE) recorded by the Beddit® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower

limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C sleep efficiency.

References

1. Cai H, Shu XO, Xiang YB, et al. Sleep duration and mortality: a prospective study of 113 138 middle-aged and elderly Chinese men and women. *Sleep* 2015;38(4):529-36. doi: 10.5665/sleep.4564 [published Online First: 2014/10/29]
2. Davona T. The Wearables Report: Growth Trends, Consumer Attitudes and Why Smart Watches Will Dominate, Business Insider, February 12, 2015. *Buisness Insider Australia* 2015 JUL 6 2015; JUL 6 2015. http://www.businessinsider.com.au/the-wearable-computing-market-report-bii-2015-7.
3. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav* 2016;158:143-9. doi: 10.1016/j.physbeh.2016.03.006 [published Online First: 2016/03/13]
4. Intelligence C. US Enterprise Wearables Market: 5-Year Forecast, 2014–2019 2015 [Available from: http://www.marketwired.com/press-release/compass-intelligence-forecasts-wearables-enterprise-grow-exponentially-us-device-revenue-2032309.htm.
5. de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents. *Sleep* 2015;38(9):1461-8. doi: 10.5665/sleep.4990 [published Online First: 2015/07/15]
6. Cook JD, Prairie ML, Plante DT. Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy. *J Clin Sleep Med* 2018;14(5):841-48. doi: 10.5664/jcsm.7120 [published Online First: 2018/05/08]
7. Kang SG, Kang JM, Ko KP, et al. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res* 2017;97:38-44. doi: 10.1016/j.jpsychores.2017.03.009 [published Online First: 2017/06/14]
8. Tuominen J, Peltola K, Saaresranta T, et al. Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. *J Clin Sleep Med* 2019;15(3):483-87. doi: 10.5664/jcsm.7682 [published Online First: 2019/03/12]
9. De Chazal P, Fox N, O'Hare E, et al. Sleep/wake measurement using a non-contact biomotion sensor. *J Sleep Res* 2011;20(2):356-66. doi: 10.1111/j.1365-2869.2010.00876.x [published Online First: 2010/08/14]
10. Zaffaroni A, de Chazal P, Heneghan C, et al. SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference*

2009;2009:7091-4. doi: 10.1109/IEMBS.2009.5332909 [published Online First: 2009/12/08]
11. Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *J Clin Sleep Med* 2012;8(5):597-619. doi: 10.5664/jcsm.2172 [published Online First: 2012/10/16]
12. AliphCom dba Jawbone. Jawbone UP3 San Francisco2016 [Available from: https://jawbone.com/support/articles/000001027/download-the-app accessed 19/4/16 2016.
13. ResMed. San Diego2016 [Available from: https://itunes.apple.com/us/app/s+-by-resmed/id883611019?mt=8 accessed 19/4/16 2016.
14. Beddit Ltd. Beddit Sleep Tracker Helsinki, Finnland2016 [Available from: http://support.beddit.com/hc/en-us/articles/201422237-Downloading-the-Beddit-app accessed 19/4/16 2016.
15. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016;15(2):155-63. doi: 10.1016/j.jcm.2016.02.012 [published Online First: 2016/03/31]
16. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* 1986;1(8476):307-10. doi: DOI 10.1016/s0140-6736(86)90837-8
17. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences* 2014;21:19-25.
18. de Zambotti M, Claudatos S, Inkelis S, et al. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiol Int* 2015;32(7):1024-8. doi: 10.3109/07420528.2015.1054395 [published Online First: 2015/07/15]
19. Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Med Rev* 2020;49:101227. doi: 10.1016/j.smrv.2019.101227 [published Online First: 2020/01/07]
20. Lee D. Apple releases new Beddit sleep trcker 2018 [Available from: https://www.theverge.com/2018/12/7/18131220/apple-beddit-3-5-sleep-monitor accessed 22.7.21 2021.
21. Dignan L. SleepScore Max review 2017 [Available from: https://www.zdnet.com/article/sleepscore-max-review-sleep-improvement-system-with-big-data-backing/ accessed 22.07.2021 2021.
22. Smith C. Rise and fall of the Jawbone UP24: The tracker that changed wearable tech 2019 [Available from: https://www.wareable.com/fitness-trackers/remembering-the-jawbone-up24-7320 accessed 22.07.2021 2021.
23. Khosla S, Deak MC, Gault D, et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *J Clin Sleep Med* 2018;14(5):877-80. doi: 10.5664/jcsm.7128 [published Online First: 2018/05/08]

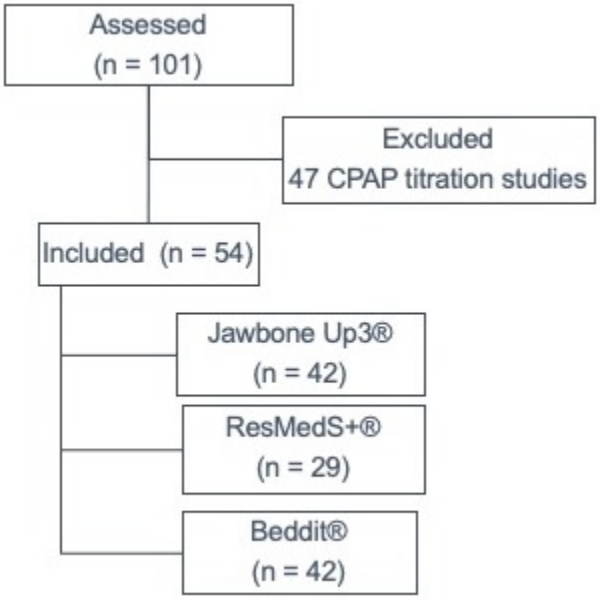


Figure 1 - Consort statement of data collection.

156x153mm (54 x 54 DPI)

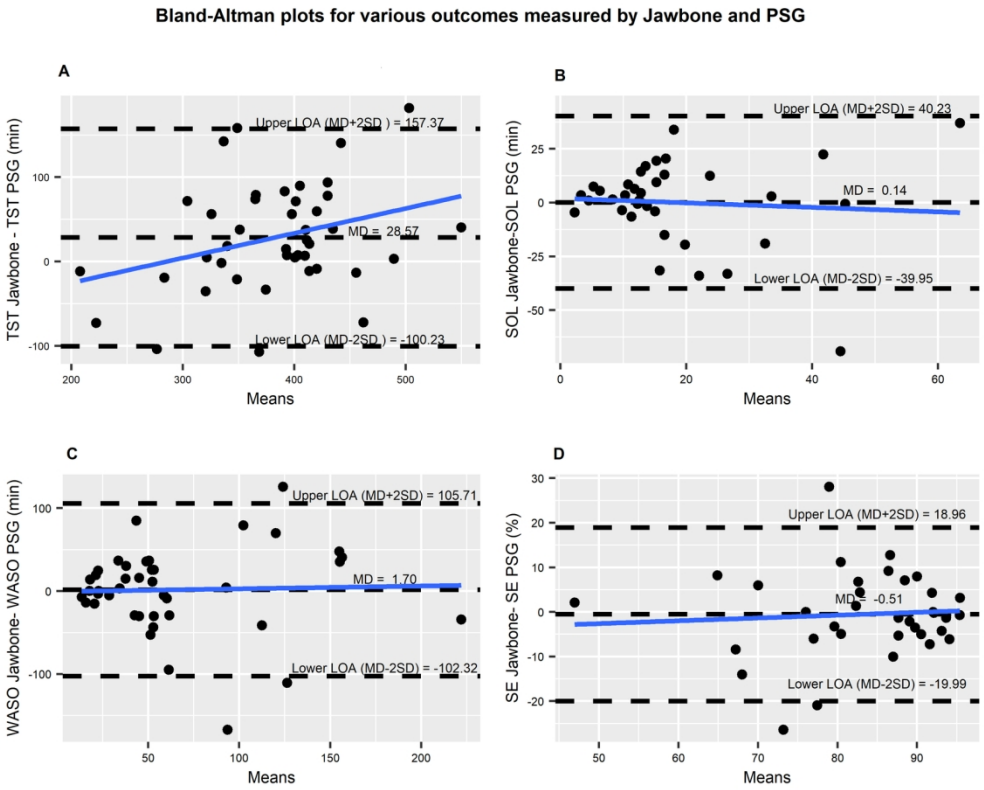


Figure 2 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement. PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

152x127mm (300 x 300 DPI)

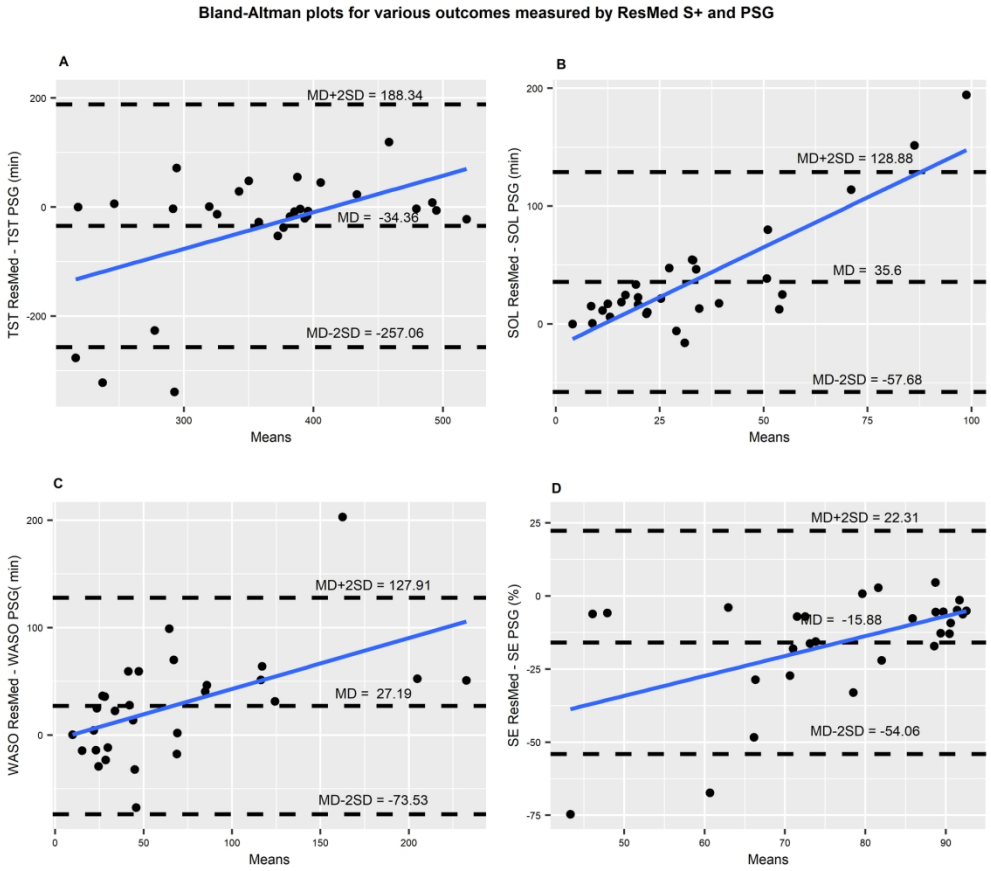


Figure 3 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the ResMed S+® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

177x165mm (300 x 300 DPI)

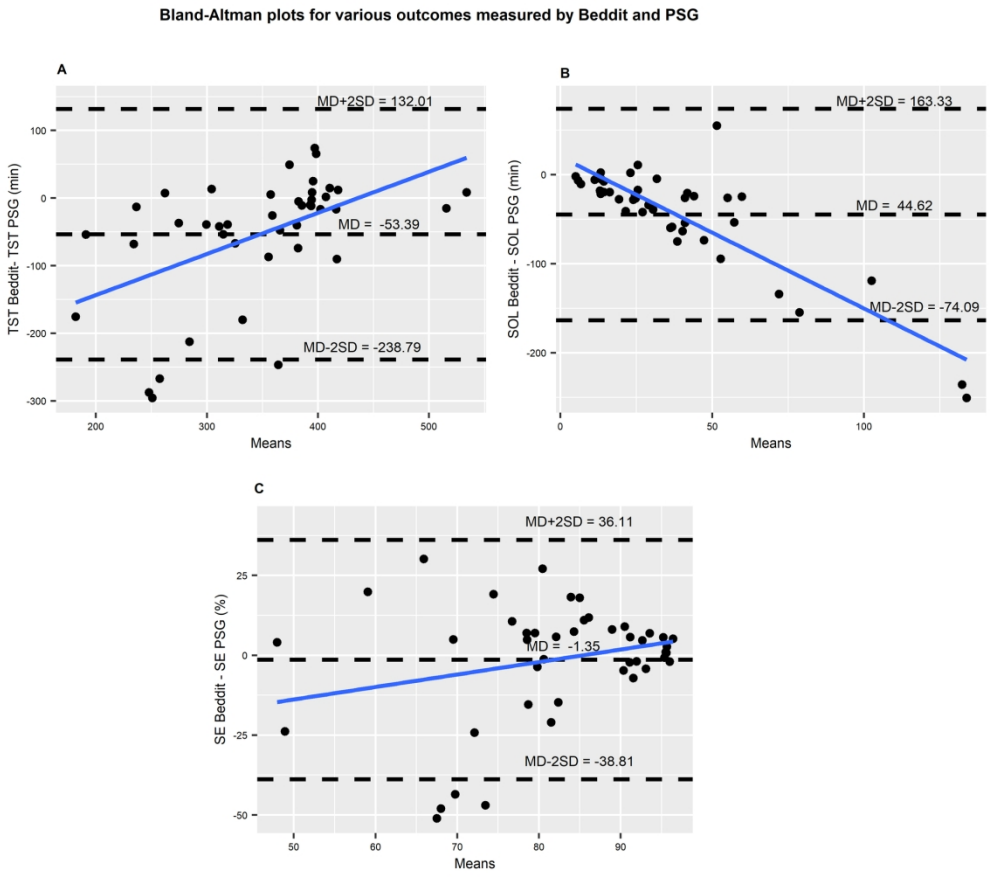


Figure 4. Bland-Altman plot of three outcomes (TST, SOL and SE) recorded by the Beddit® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C sleep efficiency.

177x165mm (300 x 300 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	1, 5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	
Participants	**6**	Eligibility criteria	5-7
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	5-7
	8	Where and when potentially eligible participants were identified (setting, location and dates)	5-7
	9	Whether participants formed a consecutive, random or convenience series	5
Test methods	**10a**	Index test, in sufficient detail to allow replication	5
	10b	Reference standard, in sufficient detail to allow replication	6
	11	Rationale for choosing the reference standard (if alternatives exist)	7
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-8
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	6
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	6
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	7-8
	15	How indeterminate index test or reference standard results were handled	7
	16	How missing data on the index test and reference standard were handled	10
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	5
	18	Intended sample size and how it was determined	-
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Figure1
	20	Baseline demographic and clinical characteristics of participants	Table 1
	21a	Distribution of severity of disease in those with the target condition	Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	-
	22	Time interval and any clinical interventions between index test and reference standard	-
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Table 2
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Table 3
	25	Any adverse events from performing the index test or the reference standard	-
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-12
	27	Implications for practice, including the intended use and clinical role of the index test	12
OTHER INFORMATION			
	28	Registration number and name of registry	-
	29	Where the full study protocol can be accessed	-
	30	Sources of funding and other support; role of funders	13

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-044015.R3
Article Type:	Original research
Date Submitted by the Author:	27-Sep-2021
Complete List of Authors:	Ellender, Claire; Princess Alexandra Hospital, ; The University of Queensland, Zahir, Syeda ; The University of Queensland, QCIF Facility for Advanced Bioinformatics Meaklim, Hailey; Melbourne Sleep Disorders Centre; St Vincent's Hospital Melbourne Pty Ltd Joyce, Rosemarie; Melbourne Sleep Disorders Centre Cunnington, David; Melbourne Sleep Disorders Centre, Swieca, John; Melbourne Sleep Disorders Centre
Primary Subject Heading:	Respiratory medicine
Secondary Subject Heading:	General practice / Family practice
Keywords:	SLEEP MEDICINE, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, RESPIRATORY MEDICINE (see Thoracic Medicine), Telemedicine < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ORIGINAL RESEARCH

Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population

Claire M. Ellender[1,2]
Syeda Farah Zahir[3]
Hailey Meaklim[1]
Rosemarie Joyce[1]
David Cunnington[1]
John Swieca[1]

[1] Melbourne Sleep Disorders Centre, East Melbourne, Victoria, Australia

[2] Princess Alexandra Hospital, Brisbane, Australia

[3] QCIF Facility for Advanced Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, 4072, AUSTRALIA

ORCID's: Hailey Meaklim 0000-0003-0448-3567; Claire Ellender 0000-0002-1727-576X; David Cunnington 0000-0002-8403-0420; Syeda Farah Zahir 0000-0002-2074-6999; John Swieca 0000-0001-8281-4048.

Correspondence: Dr Claire Ellender
Department of Respiratory & Sleep Medicine, Princess Alexandra Hospital, 199 Ipswich Rd, Brisbane, Qld, Australia 4170
Tel +61731762698, Fax +61 731766170
Email Claire.ellender@health.qld.gov.au

Word Count: 3180

Keywords: Consumer sleep monitor; accelerometer, polysomnography; sleep; sleep measurement; validation; wearables

ABSTRACT:

Objectives: Consumer-grade smart devices are now commonly used by the public to measure waking activity and sleep. However, the ability of these devices to accurately measure sleep in clinical populations warrants more examination. The aim of the present study was to assess the accuracy of three consumer-grade sleep monitors compared with gold standard polysomnography (PSG).

Design: A prospective cohort study was performed.

Setting: Adults undergoing PSG for investigation of a suspected sleep disorder.

Participants: 54 sleep-clinic patients were assessed using three consumer-grade sleep monitors (Jawbone UP3®, ResMed S+® and Beddit®) in addition to PSG.

Outcomes: Jawbone UP3®, ResMed S+® and Beddit® were compared with gold standard in-laboratory polysomnography on 4 major sleep parameters - total sleep time (TST), sleep onset latency (SOL), Wake After Sleep Onset (WASO) and sleep efficiency (SE).

Results: The accelerometer Jawbone UP3® was found to overestimate TST by 28mins (limits of agreement, LOA= -100.23 to 157.37), with reasonable agreement compared with gold standard for TST, WASO and SE. The doppler radar ResMed S+® device underestimated TST by 34mins (LOA= -257.06 to 188.34) and had poor absolute agreement compared with PSG for TST, SOL and SE. The mattress device, Beddit® underestimated TST by 53mins (LOA= -238.79 to 132) on average and poor reliability compared to PSG for all measures except TST. High device synchronisation failure occurred, with 20% of recordings incomplete due to Bluetooth drop out and recording loss.

Conclusion: Poor to moderate agreement was found between PSG and each of the tested devices, however Jawbone UP3® had relatively better absolute agreement than other devices in sleep measurements compared with PSG. Consumer grade devices assessed do not have strong enough agreement with gold standard measurement to replace clinical evaluation and PSG sleep testing. The models tested here have been superseded and newer models may have increase accuracy and thus potentially powerful patient engagement tools for long-term sleep measurement.

Strengths and Limitations of this study

- Consumer grade devices were compared with gold standard in clinic patients.
- More than one device was included for comparison.
- This study includes measure of sleep parameters that clinicians frequently need to review in daily practice, such as total sleep time and sleep efficiency.
- High device failure was found in this study, confirming that consumer grade devices cannot be used to replace high fidelity diagnostic measurement.
- This sample had patients with sleep apnoea, insomnia or hypersomnia as their final sleep diagnosis.

BACKGROUND

Poor sleep quality and duration has been shown to be an independent risk to overall mortality and for many chronic diseases.[1] The gold standard test for the measurement of sleep and diagnosis of sleep disorders is attended polysomnography (PSG). However, this is an involved and costly test, that requires complex equipment, dedicated space, trained staff, and does not lend itself well to multi-night monitoring.

Sales of consumer sleep monitors and wearable consumer-grade smart devices have dramatically increased in recent years, with 33 million units estimated to have been sold in the United States in 2015[2] and the estimated value of the wearable industry in the USA expected to grow to $USD8.5 billion in 2020.[3 4] Consumer-grade devices fall into three major categories (i) wrist based devices (eg Jawbone®, FitBit®); (ii) Bedside devices (eg ResMed S+®, Touch-Free Life Care®); and (iii) Mattress-based devices (eg Beddit®, EarlySense Mattress®, Emfit Bed Sensor®). Each of the categories of devices utilise unique proprietary algorithms for inferring wake/sleep, body position and measures of sleep quality.

The Jawbone UP® (the precursor to the UP3 used in this study) has been compared to PSG in adolescents and concluded to have good agreements for TST, SE and wake after sleep onset (WASO), however the tendency to underestimate TST and sleep efficiency increased with age.[5] In a study of adult women, the FitBitChargeHR® overestimated TST by 27min, and was found to have significantly different SOL and WASO compared to PSG.[5] Similarly in adolescents the Jawbone UP® tended to overestimate TST and SOL, whilst underestimating WASO. The researchers also found greater discrepancies in nights when participants had more disrupted sleep (i.e., lower TST and greater SOL and WASO).[5] In patients with suspected central disorders of hypersomnolence, the Jawbone UP3® was found to significantly overestimate TST by an average of 39.6 minutes compared to PSG and was not able to discriminate stages of sleep adequately.[6] Interestingly, the Jawbone UP3® performed similarly to actigraphy in this study. Another clinical study found that the FitBit Flex® overestimated TST more in a group of insomnia patients compared to good sleepers (32.9 mins vs. 6.5 mins).[7] Taken together, these two studies suggest that consumer-grade sleep devices are less accurate at measuring TST in a clinical sleep disorder population, than they are for good sleepers.

The Beddit® mattresses based device has been found in 10 health controls to have poor agreement with TST (overestimated by 43.5min), WASO and SE.[8] SOL was the only measure to have agreement, but had a wide variance.[8] The sensor technology used in the ResMed S+® device has been shown to have moderate accuracy in measuring TST and sleep efficiency in

healthy volunteers compared to PSG and high specificity.[9 10] Furthermore its utility in measuring sleep disordered breathing has been investigated and found to have reasonable accuracy in detecting moderate obstructive sleep apnoea, with a sensitivity of 89% and specificity 92%.[11]

Patients are increasingly attending sleep clinics with downloads from consumer-grade devices for discussion with primary care physicians and sleep specialists. These commonly encountered situations in the sleep clinic raise the questions: how reliable are consumer-grade devices, and which type of technology is most comparable to gold standard? This study aims to answer these questions with an in-laboratory comparison of PSG with the three consumer devices - Jaw Bone UP3®, Beddit® and ResMed S+® in a sleep clinic population. It was hypothesized that these devices would have similar accuracy in detecting TST, SOL, WASO and SE.

METHODS

Study Population

54 adult patients were consecutively recruited through a private sleep disorders centre in Melbourne, Australia from June 2015 to February 2016. Inclusion criteria were age >18years and any patient who required overnight polysomnography as standard investigation following sleep physician review to either confirm or exclude sleep disordered breathing. All patients attending the laboratory for a polysomnogram were screened for inclusion. Exclusion criteria were age <18years, positive airway pressure titration study, pregnancy and cognitive impairment. Figure 1 demonstrates the consort statement.

Procedure

All assessments took place at an attended sleep laboratory in Melbourne, Australia. The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15). Sleep laboratory staff were trained to set up the 3 devices in addition to regular overnight polysomnography monitoring; lights out time was noted for synchronisation across all devices. The primary outcome measure was Total Sleep Time (TST) and secondary outcomes were sleep onset latency (SOL, min), sleep efficiency (SE, %) as TST/(TST+ total wake time) and wake after sleep onset (WASO, min). Other measures from the consumer grade devices such as time spent in light, deep or rapid eye movement sleep was not compared in this analysis.

Polysomnographic Recording

PSG was measured using a standard six-channel electroencephalography, submental electromyography and electrooculography, electrocardiogram, airflow (thermistor & nasal cannula), respiratory effort, oximetry, snoring (dB sound meter), body position, pulse rate, leg electromyography and digital video, recorded according to American Academy of Sleep Medicine standards.[12] The following standard sleep parameters were recorded via PSG: Total sleep time (TST), sleep onset latency (SOL, min), total wake time (TWT, min), sleep efficiency (SE, %) as TST/(TST+TWT) and wake after sleep onset (WASO, min). Participants were classified as having obstructive sleep apnoea if the apnoea hypopnoea index was >5 events/hr. A single registered polysomnographic technologist scoring the PSG was blinded to the download of consumer grade devices and raw data was scored using Compumedics® amplifiers and Profusion software version 3 (Compumedics®, Abbotsford, Victoria, Australia).

JawBone UP3®

Participants were fitted with the JawBone Up3® on the participant's non dominant wrist with the Jawbone UP3® shortly before lights out time. Data was collected via a dedicated iPod Touch, synced to the Jawbone® app version 4.0.0.[13] This consumer-grade actigraphy device has a three-axis accelerometer and heart rate monitor, which together measure TST, SOL, WASO and SE which were exported by a technician the following morning after the PSG was complete.

ResMed S+®

The ResMed S+® is a non-contact radio-frequency sensor that continuously measures the biomotion due to breathing and body-movement in bed. The sensor operates in a license-free band at 5.8 GHz, emits an average power less than 1 mV and is capable of sensing movement and breathing over a distance ranging from 0.3 to 1.5 meters. The device was positioned by the bedside and synced shortly before lights out time to a dedicated iPod with the ResMed S+® app Version 1.2.1. [14] Measurements from the ResMed S+® were TST, SOL, WASO, SE which were exported by a technician the following morning after the PSG was complete.

Beddit®

The primary sensor in the Beddit® is a piezoelectric 70cm band that was attached to the mattress prior to patients getting into bed. The device detects micro-movements of the chest wall from heartbeats and respiration and uses ballistocardiography to infer sleep stage and time. Ballistocardiography is a non-invasive measurement of cardiac output and respiration by

converting mechanical motion (e.g. movement generated by a heartbeat) to a digital signal. Measurements from the Beddit® were taken each night using the device synced to a dedicated iPod running the Beddit® app version 1.[15] Output from the app included TST, SOL, WASO, SE and HR which were exported by a technician the following morning after the PSG was complete.

Statistical analyses

Each of the three non-invasive devices were compared with PSG as the gold standard on an intention to treat basis. The primary and secondary outcomes were compared on total measurements over the night, not epoch-by-epoch method. Summary statistics of the study population are presented. For all normally distributed continuous variables mean and SD, whereas for non-normally distributed variables median and IQR were presented. Normality was assessed using the Shapiro-Wilk's test. Frequencies and proportions are presented for categorical variables. Extent of agreement and reliability between gold standard and each of the selected test devices, was assessed using Intraclass Correlation Coefficients (ICCs) with two-way random-effects model. Agreement was considered moderate, good and excellent if the ICC values were between 0.5 and 0.75, 0.75 and 0.9 and >0.9 respectively[16].

Additionally, Bland-Altman plots[17] were used to visualize the agreement between gold standard polysomnography and each of the selected devices. The average of two measurements was plotted on x-axis and difference between the two along y-axis. The mean of the differences provided an estimate of average bias between the methods. The upper and lower Limits of agreement (LOA) were calculated which correspond to the mean difference (Gold standard–Selected method) ±2 standard deviations (SD). LOA estimated the interval that a given proportion of differences between the measurements is likely to lie within and will be used to determine if the methods can be used interchangeably. Cohen's d is reported for the magnitude of the effect size. In case of non-normally distributed data, effect size 'r' was calculated by dividing Z statistic by the square root of the sample size (N)). Interpretation of r is 0.10 - < 0.3 (small effect), 0.30 - < 0.5 (moderate effect) and >= 0.5 (large effect)[18]. Data were analysed using R (4.0.4) (https://www.r-project.org/) (R Core Team, 2017).

Patient and Public involvement

Patients at our sleep disorders centre sparked the interest to assess the accuracy of consumer-grade sleep monitors. Our clinicians were often asked about the accuracy of home sleep monitors. To answer this question our team invited the patients to be involved in

evaluating three commonly available consumer-grade smart devices. Participants were not paid for their involvement but did provide written consent. The findings of this research suggest that consumer-grade sleep monitors can give insights into trends in sleep but are not accurate enough to replace laboratory measurement.

RESULTS

Fifty-four adult patients (57% females) with a mean age of 48.09 (±SD 18.05) years participated in this study. Table 1 presents demographics of study population. The final sleep diagnosis found was obstructive sleep apnoea in 33 (61%), insomnia 9 (17%) and central hypersomnolence disorder in 12 (22%) participants. The mean PSG detected TST was 371min (SD±69), SOL of 16min (SD±15), WASO 63min (SD±56) and SE of 82% (SD±13%). The absolute values of the measurements for each device are summarised in table 2. The results of the Bland-Altman analyses and intra-class correlation are summarized in table 3 and displayed in figures 2-4.

JawBone UP3®

On average JawBone UP3® overestimated TST by 28.57mins (LOA= -100.23 to 157.37). By inspecting the Bland-Altman plots (shown in Figure 2A), the cluster of points surrounded the mean tightly between 300-400 minutes and there was greater variability with TST below 300min and above 400 minutes. The magnitude of effect size was small (d=0.44). A moderate degree of reliability for recording TST was found between PSG and Jawbone UP3® with an intra-class correlation coefficient (ICC) of 0.6 (95% CI = 0.34 to 0.77; p<0.001).

Bland Altman plot (Figure 2B) suggests that the mean difference in SOL between two methods was very small and on average JawBone UP3® measured SOL 0.14mins (LOA= -39.95 to 40.23) more than the gold standard. The cluster of points surrounded the mean tightly on the left, with greater variability for values over 20 minutes. The magnitude of difference was small (r=0.13). The reliability between the two methods was between poor to moderate (ICC=0.29; 95% CI=-0.04 to 0.57; p=0.04).

Jawbone UP3® overestimated WASO only slightly, 1.7min (LOA= -102.32 to 105.71, d=0.03) compared with PSG. Greater variability was seen for measurements over 50 minutes (as shown in Figure 2C), indicating better estimation of WASO by JawBone UP3® at lower values. The agreement between Jawbone UP3® and PSG for WASO was poor to moderate (ICC=0.55; 95% CI= 0.29-0.73; p <0.001).

The mean difference in SE between two methods indicated that on an average JawBone UP3® measures SE 0.51% (LOA: -18.96 to 19.99) less than the gold standard. This bias seems to be due to measurements less than 85%, with better estimation of SE by JawBone UP3® at higher SE, as seen in Figure 2D. The magnitude of difference was small (d=0.05) The ICC for agreement between Jawbone Up3® and PSG regarding SE was 0.66 (95% CI=0.41 to 0.81; p<0.001) indicating poor to good reliability between the two measures based on 95% CI.

ResMed S+®

As shown in Figure 3A, on average ResMed S+® underestimated TST by 34 minutes (CI: -257 min to 188 min). The mean difference between ResMed S+® measured and PSG measured TST was offset (lying below) zero, suggesting a bias. The points remained in the same general pattern for all x-axis values, except for few outliers at lower mean values. The magnitude of difference was moderate (r=0.4). ICC of 0.36 (95% CI: 0.02-0.63; p=0.02) indicating poor to moderate reliability.

Conversely, ResMed S+® overestimated SOL by 35.6 min (LOA = -57.68 to -128.89) and effect size was large (r=0.8). Cluster of points go from below the mean at short SOL, to above the mean with increasing SOL, showing proportional error, suggesting overestimation of SOL by ResMed S+® at increasing SOL duration, as shown in Figure 3B. A poor agreement for SOL was seen between the two methods (ICC= -0.01; 95% CI: -0.21 – 0.26; p=0.51).

Similarly, ResMed S+® recorded WASO 27 min more than PSG (LOA= -73.53 to 127.91) and a large effect was found (r=0.52). Visual inspection of Bland Altman plot (Figure 3C) suggested that ResMed S+® increasingly overestimating WASO with increasing time. Reliability between methods was between poor to excellent (ICC= 0.61; 95% CI= 0.28 to 0.8, p <0.01).

Visual inspection of the Bland-Altman plot Figure 3D suggests that on average ResMed S+® underestimated SE by 16% (LOA=-54.06 to 22.31). The effect size was large (r=0.8) and an ICC value of 0.28 (95% CI= -0.06 to 0.58; p=0.06) was found. Moreover, the mean difference was not constant, with greater variability at lower values (particularly below 80%), showing proportional bias.

Beddit®

The Beddit® and PSG had the least agreement for all outcomes except TST compared to other devices. TST was underestimated by 53min (LOA= -238.79 to 132). As demonstrated in Figure 4A, the cluster of points shifted from below mean to above mean with increasing TST,

showing a proportional error depending on the duration of sleep. The magnitude of difference was large (r=0.55) and reliability poor to moderate (ICC= 0.40; 95% CI=0.09 to 0.63; p =0.01).

SOL was overestimated by 45min (LOA= -74.09 to 163.33) by the Beddit® compared with PSG. The points were tightly clustering above the mean, and go from above, to below the mean, from left to right (Figure 4B), showing error proportional to the duration of SOL. The effect size was large (r=0.78) and reliability poor (ICC = 0.004; 95% CI=-0.173 to 0.22; p=0.48).

Beddit® slightly underestimated SE by 1.35% (LOA= -38.81 to 36.11). As shown in Figure 4C, variability of points was constant around the mean at values below 80%. This suggest that at higher values, Beddit® estimated SE more closely to the PSG gold standard. The effect size was small (r=0.13) and poor agreement (ICC 0.26; 95% CI=-0.04 to 0.51; p=0.06).

Consumer-grade recording failure

Consumer-grade devices were set-up by Sleep Scientist staff each night at the time of the standard PSG set-up. Despite this, device or recording failure resulting in inability to record sufficient data, on the single night of recording, in the consumer-grade devices was common. Failure to synchronise with the dedicated Bluetooth device was the most common reason for device failure. The ResMed S+® failed to synchronise the most, with 25/54 nights (46%) resulting in recording failure. The Jawbone and Beddit® had similar rates of synchronisation failure (12/54, 22%), however not usually in the same room or on the same patient. Comparisons were made on an intention to treat analysis, even where large differences in TST were seen.

DISCUSSION

The agreement of these three consumer-grade smart devices have simultaneously been compared with gold standard attended PSG in an adult sleep clinic cohort. For each of the devices, there were components of sleep measurement with poor to moderate agreement with the gold standard. This study found the primary outcome measure of TST was overestimated by, Jawbone UP3® whereas both ResMed S+® and Beddit® underestimated it. The Jawbone UP3® also overestimated SOL and WASO, however the magnitude of difference was very small. Generally Jawbone UP3® had better agreement across all outcomes, however for SE

agreement was better between ResMed S+® and PSG. The Beddit® had the least agreement with PSG, all components having poor agreement when compared with gold standard PSG.

Wearable devices, particularly wrist-worn accelerometers have now been widely compared with PSG. Similar to the results of this study, the accelerometers have been shown to overestimate total sleep time by around 20-30minutes, particularly in sleep disordered populations compared with healthy controls.[5 7 19] Previous investigations into consumer grade accelerometers in clinical populations found TST overestimated by 32.9min[7] in a population of 33 insomnia patients and 39min in 43 hyper-somnolence patients[6]. In our study SOL had a large confidence interval, with bias found with measurements over 15min, consistent with findings of a recent systematic review and meta-analysis.[20]

The Beddit® device and mattress devices in general are one of the least studied consumer grade devices. Tuominen *et al.* (2019) found in 10 healthy controls the Beddit® overestimated total sleep time by 43min, whereas our data suggests a significant underestimation (PSG TST 371min versus Beddit® TST 321 min) with a larger sample size (n = 42). Tuominen *et al.* (2019) was also able to access WASO data, which was not available with the model of Beddit® tested in this study and found to underestimate WASO by 32min. Non-wearable devices have a potential growing market as non-intrusive home monitors of sleep, as they can be applied in a "set and forget" method. Thus, further refinement and evaluation of bed-based devices would be desirable.

Chinoy *et al.* (2021)[10] recently compared PSG to ResMed S+® and to SleepScore Max with a population of 19 young 'healthy normal' individuals. The ResMed S+® was found to have underestimated TST by only 0.3min (95%CI: -70.7-70.2) and the SleepScore Max overestimate TST by 7.5min (-60.7 to 75.7). A likely explanation for the difference these findings and the present study is the difference in population – 'healthy normal' participants versus sleep clinic population. There is growing literature that consumer grade devices have lower accuracy in clinical population compared with control populations.[21] Notably, Chinoy *et al.* (2021)[10] found 2/19 nights (10.5%) using the ResMed S+® were impacted by device synchronisation issues, requiring device re-synchronisation.

The high device synchronisation failure rate also observed in our study is concerning, despite the set-up being performed by sleep laboratory scientific staff. There is no way to calibrate these consumer-grade devices over time and it is difficult to monitor device connectivity to the Bluetooth device until the next morning. The high failure rate further confirms the role of these consumer devices is not to replace that of a diagnostic sleep study.

The main strength of this study was the sample size and that it was conducted in a clinical adult sleep population with a range of suspected sleep disorders. This makes the findings more translatable to clinicians managing patients with sleep disorders. Further, assessing a number of different devices is a novel approach. The weaknesses of the study include a high device recording failure rate, predominantly with Bluetooth synchronisation failure. Epoch-by-epoch analysis was not performed. Further, sales of devices tested in this study have since been discontinued. Beddit® was acquired by Apple Inc in May 2017 and relaunched an updated device, the Beddit® 3.5 which has reportedly improved integration with mobile phone health kits[22]. The ResMed S+® was discontinued and subsequently a similar device was launched in 2017 as SleepScore labs, which is similarly Apple iOS and Adroid integrated[23]. JawBone® however has gone into liquidation with no subsequent models leading on from the UP3® device[24].

This study indicates that the wrist worn Jawbone UP3® had the best agreement in measuring sleep compared with gold standard and can provide useful information about commonly measured parameters of sleep quality. For Sleep Medicine Clinicians, the translation of these findings, is that when our patients present with longitudinal measurements of sleep from their consumer grade devices, we can be reassured that wrist worn devices have reasonably accuracy and can be harnessed as an engagement tool for behavioural sleep interventions. This is consistent message with the American Academy of Sleep Medicine's position statement about the use of consumer-grade sleep devices stating that these devices cannot be used for clinical diagnosis, however they allow for meaningful discussions with patients about sleep and encourage active participation in sleep-related health care.[25]

CONCLUSION

Given the large body of literature linking sleep quality to mortality and many chronic diseases, patient-collected longitudinal sleep data provides a powerful insight into a patient's overall health. This study adds to the data of consumer grade wearable sleep monitors, showing they can provide some reliable information compared to gold standard PSG, however do not replace clinical evaluation and gold-standard PSG sleep testing. In reviewing sleep data collected by patients with consumer-grade devices, clinicians are encouraging measurement and quantification of sleep, which in turn will likely emphasise the importance of quality sleep in maintaining good health.

ACKNOWLEDGMENTS

Sleep laboratory staff at St Vincent's Private Hospital, East Melbourne for their set up efforts. Telstra Corporation Ltd (Australia) for the provisions of the Jawbone UP3, ResMed (San Diego) for the ResMed S+ and Beddit Ltd (Finland) for the supply of the test devices used. The authors acknowledge the statistical support received through the Metro South Health Biostatistics Service.

Ethics Approval

The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15).

Dataset availability

The dataset will be available upon emailed request to the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public or not-for-profit sectors.

Competing Interests

The Telstra Corporation Ltd (Australia) provided the Jawbone UP3 test devices used in the study, ResMed (San Diego) provided the ResMed S+ and Beddit Ltd (Finland) provided the Beddit device.

Author contributions

Dr Claire M. Ellender was involved in the protocol preparation, participant consent, data collection, analysis and manuscript preparation. Dr Syeda Farah Zahir was involved in the data curation and analysis and manuscript preparation. Ms Hailey Meaklim was involved in data analysis and manuscript preparation. Ms Rosemarie Joyce was involved with participant consent, data collection and manuscript preparation. Dr David Cunnington and Dr John Swieca were involved in protocol preparation, data analysis and manuscript preparation.

Table 1. Patient demographics

Variable	Results (n = 54)
Age in years, mean (SD)	48.09 (±SD 18.05)
Gender	31 (57%) women
	23 (43%) men
BMI kg/m^2, median (IQR)	27 (24-31)
PSG AHI events/hr, median (IQR)	9 (3-18.75)
Indication for PSG	
Rule in suspected OSA	32 (60%)
Rule out OSA	22 (40%)
Final clinical diagnosis	
OSA syndrome	33 (61%)
Insomnia	9 (17%)
Hypersomnia	12 (22%)

PSG, Polysomnogram; BMI, Body Mass Index; AHI, Apnoea hypopnoea index; OSA, Obstructive sleep apnoea

Table 2 Mean sleep duration

		DEVICE		
VARIABLE	PSG	Jawbone UP3® (N = 42)	ResMed S+® (N = 29)	Beddit® (N = 42)
TST (MIN SD±)	371 ±69	397 ±83	345.8 ±120	321 ±107
SOL (MIN)	16 ±15	18 ±16	50 ±44	60 ±57
WASO (MIN)	63 ±56	65 ±55	80 ±72	-
SE (%)	82.4 ±13	82.9 ±11	68.8 ±21	81 ±17

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency.

Table 3. Comparison of the outcomes between polysomnography (gold standard) and each of the selected methods

	TST (min)	SOL (min)	WASO (min)	(%)
Jawbone vs PSG Bland-Altman Analysis				
N	42	36	41	35
Bias	28.57	0.14	1.70	-0.51
LOA	-100.23 to 157.37	-39.95 to 40.23	-102.32 to 105.71	- 19.99 to 18.96
Cohen's d or r (Magnitude)	0.44 (Small)	0.13* (Small)	0.03 (Small)	0.05 (Small)
ICC	0.6 (95% CI= 0.34-0.77; p<0.001)	0.29 (95% CI= -0.04-0.57; p=0.04)	0.55 (95% CI= 0.29-0.73; p <0.001).	0.65 (95% CI=0.41-0.81; p<0.001)
ResMed S+ vs PSG Bland-Altman Analysis				
N	29	29	29	29
Bias	-34.36	35.60	27.19	-15.88
LOA	-257.06 to 188.34	-57.68 to -128.89	-73.53 to 127.91	-54.06 to 22.31
Cohen's d or r (Magnitude)	*0.41 (Moderate)	*0.81 (Large)	*0.52 (Large)	*0.8(Large)
ICC	0.36 (95% CI: 0.02-0.63; p=0.02)	-0.01 (95% CI:- 0.21-0.26; p=0.51)	0.61 (95% CI= 0.28-0.8; p <0.01)	0.06 (95% CI= -0.06-0.58; p=0.06)
Beddit vs PSG. Bland-Altman Analysis				
N	42	42	NA	44
Bias	-53.39	44.62	NA	-1.35
LOA	-238.79 to 132	-74.09 to 163.33	NA	-38.81 to 36.11
Cohen's d or r (Magnitude)	*0.55(Large)	*0.78(Large)	NA	*0.31 (Small)

ICC	0.40 (95% CI=0.09-0.63; p =0.01)	0.004 (95% CI=-0.173-0.22; p=0.48)	NA	0.26;95% CI=-0.04 to 0.51; p=0.06

PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. N=count of pairwise complete cases in groups; LOA=Limits of Agreement (MD±2SD) * effect size=r. Bias = the mean differences between test device minus PSG.

Figure captions

Figure 1 CONSORT statement of included participants. CPAP, Continuous Positive Airway pressure.

Figure 2 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

Figure 3 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the ResMed S+® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

Figure 4. Bland-Altman plot of three outcomes (TST, SOL and SE) recorded by the Beddit® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The

blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C sleep efficiency.

References

1. Cai H, Shu XO, Xiang YB, et al. Sleep duration and mortality: a prospective study of 113 138 middle-aged and elderly Chinese men and women. *Sleep* 2015;38(4):529-36. doi: 10.5665/sleep.4564 [published Online First: 2014/10/29]
2. Davona T. The Wearables Report: Growth Trends, Consumer Attitudes and Why Smart Watches Will Dominate, Business Insider, February 12, 2015. *Buisness Insider Australia* 2015 JUL 6 2015; JUL 6 2015. http://www.businessinsider.com.au/the-wearable-computing-market-report-bii-2015-7.
3. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav* 2016;158:143-9. doi: 10.1016/j.physbeh.2016.03.006 [published Online First: 2016/03/13]
4. Intelligence C. US Enterprise Wearables Market: 5-Year Forecast, 2014–2019 2015 [Available from: http://www.marketwired.com/press-release/compass-intelligence-forecasts-wearables-enterprise-grow-exponentially-us-device-revenue-2032309.htm.
5. de Zambotti M, Claudatos S, Inkelis S, et al. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiol Int* 2015;32(7):1024-8. doi: 10.3109/07420528.2015.1054395 [published Online First: 2015/07/15]
6. Cook JD, Prairie ML, Plante DT. Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy. *J Clin Sleep Med* 2018;14(5):841-48. doi: 10.5664/jcsm.7120 [published Online First: 2018/05/08]
7. Kang SG, Kang JM, Ko KP, et al. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res* 2017;97:38-44. doi: 10.1016/j.jpsychores.2017.03.009 [published Online First: 2017/06/14]
8. Tuominen J, Peltola K, Saaresranta T, et al. Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. *J Clin Sleep Med* 2019;15(3):483-87. doi: 10.5664/jcsm.7682 [published Online First: 2019/03/12]
9. De Chazal P, Fox N, O'Hare E, et al. Sleep/wake measurement using a non-contact biomotion sensor. *J Sleep Res* 2011;20(2):356-66. doi: 10.1111/j.1365-2869.2010.00876.x [published Online First: 2010/08/14]

10. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* 2021;44(5) doi: 10.1093/sleep/zsaa291 [published Online First: 2020/12/31]
11. Zaffaroni A, de Chazal P, Heneghan C, et al. SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference* 2009;2009:7091-4. doi: 10.1109/IEMBS.2009.5332909 [published Online First: 2009/12/08]
12. Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *J Clin Sleep Med* 2012;8(5):597-619. doi: 10.5664/jcsm.2172 [published Online First: 2012/10/16]
13. AliphCom dba Jawbone. Jawbone UP3 San Francisco2016 [Available from: https://jawbone.com/support/articles/000001027/download-the-app accessed 19/4/16 2016.
14. ResMed. San Diego2016 [Available from: https://itunes.apple.com/us/app/s+-by-resmed/id883611019?mt=8 accessed 19/4/16 2016.
15. Beddit Ltd. Beddit Sleep Tracker Helsinki, Finnland2016 [Available from: http://support.beddit.com/hc/en-us/articles/201422237-Downloading-the-Beddit-app accessed 19/4/16 2016.
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016;15(2):155-63. doi: 10.1016/j.jcm.2016.02.012 [published Online First: 2016/03/31]
17. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* 1986;1(8476):307-10. doi: DOI 10.1016/s0140-6736(86)90837-8
18. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences* 2014;21:19-25.
19. de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents. *Sleep* 2015;38(9):1461-8. doi: 10.5665/sleep.4990 [published Online First: 2015/07/15]
20. Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Med Rev* 2020;49:101227. doi: 10.1016/j.smrv.2019.101227 [published Online First: 2020/01/07]
21. Kahawage P, Jumabhoy R, Hamill K, et al. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: In-lab validation against polysomnography. *J Sleep Res* 2020;29(1):e12931. doi: 10.1111/jsr.12931 [published Online First: 2019/10/19]
22. Lee D. Apple releases new Beddit sleep trcker 2018 [Available from: https://www.theverge.com/2018/12/7/18131220/apple-beddit-3-5-sleep-monitor accessed 22.7.21 2021.

23. Dignan L. SleepScore Max review 2017 [Available from: https://www.zdnet.com/article/sleepscore-max-review-sleep-improvement-system-with-big-data-backing/ accessed 22.07.2021 2021.
24. Smith C. Rise and fall of the Jawbone UP24: The tracker that changed wearable tech 2019 [Available from: https://www.wareable.com/fitness-trackers/remembering-the-jawbone-up24-7320 accessed 22.07.2021 2021.
25. Khosla S, Deak MC, Gault D, et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *J Clin Sleep Med* 2018;14(5):877-80. doi: 10.5664/jcsm.7128 [published Online First: 2018/05/08]

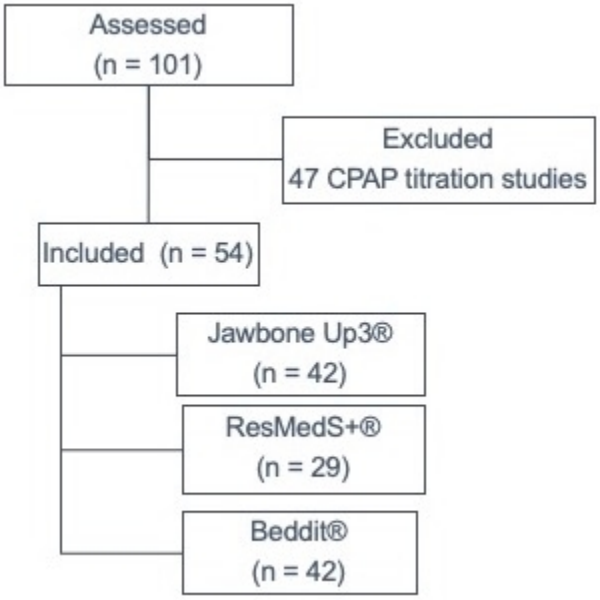


Figure 1 - Consort statement of data collection.

156x153mm (54 x 54 DPI)

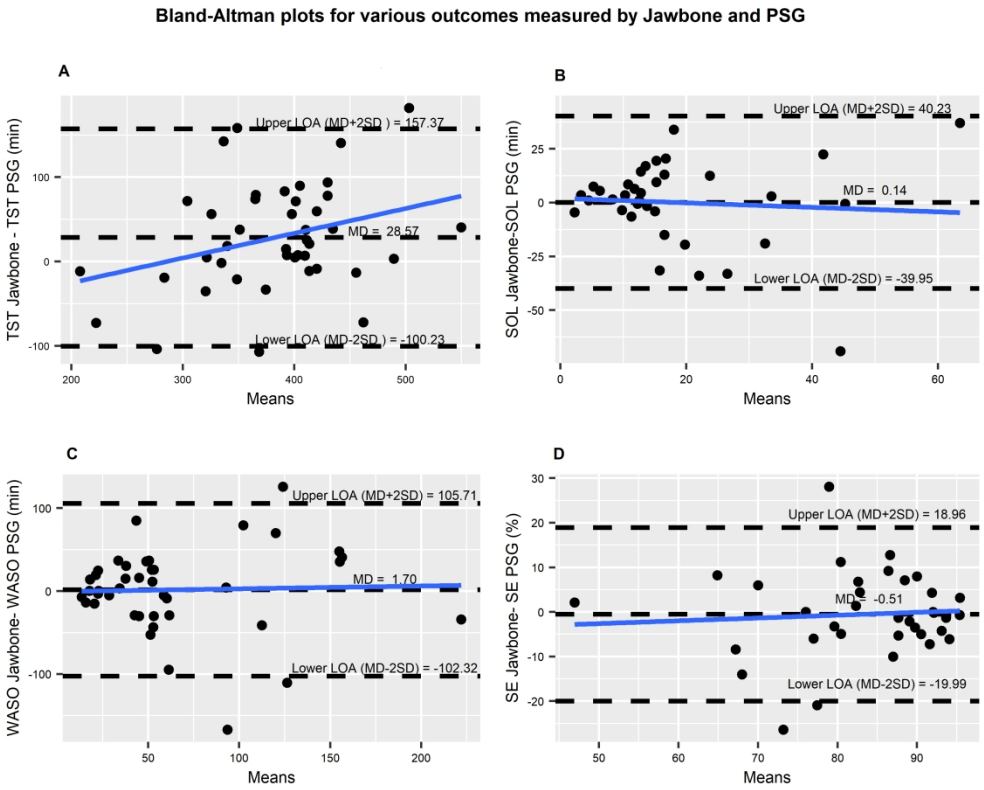


Figure 2 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the Jawbone UP3® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement. PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

152x127mm (600 x 600 DPI)

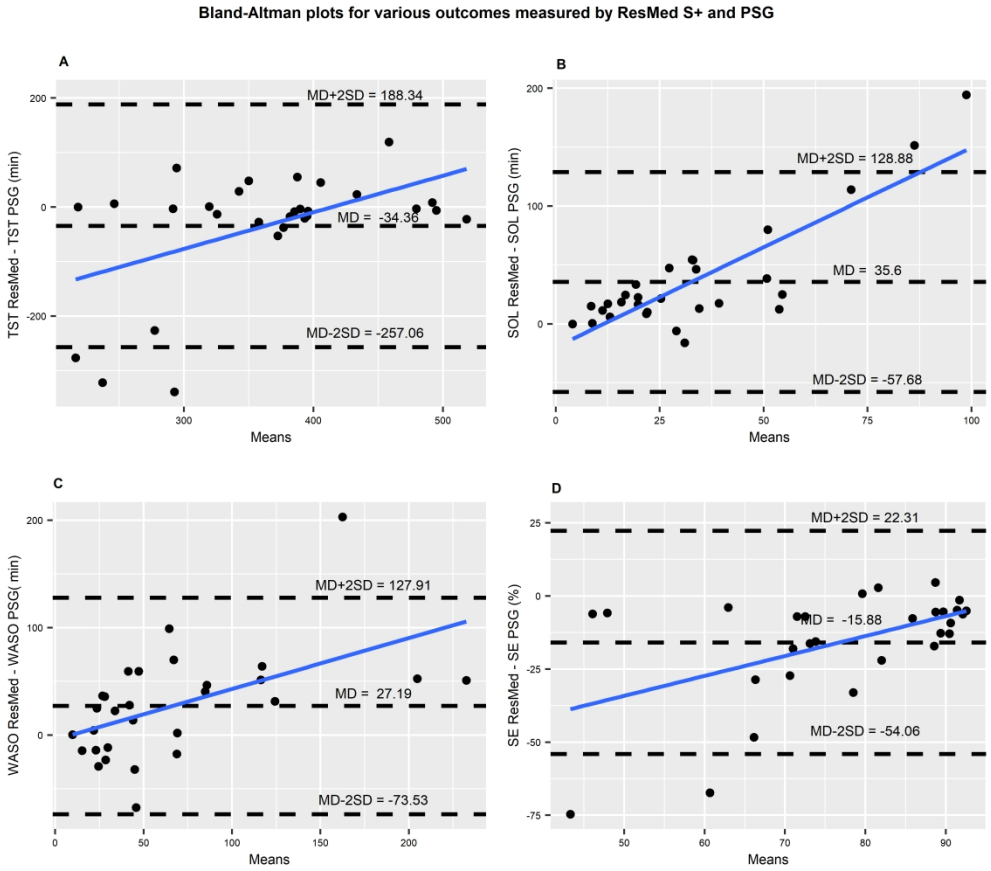


Figure 3 Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the ResMed S+® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency; WASO, Wake After Sleep Onset; and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C wake after sleep onset and Panel D sleep efficiency.

177x165mm (600 x 600 DPI)

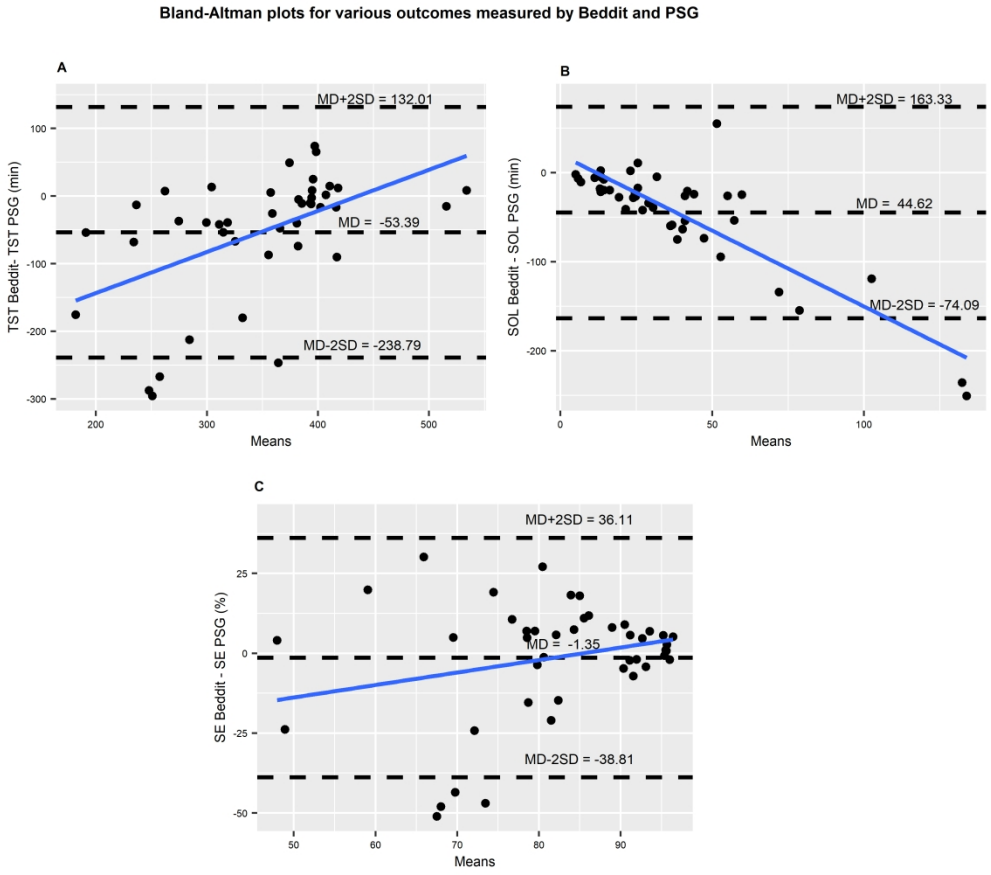


Figure 4. Bland-Altman plot of three outcomes (TST, SOL and SE) recorded by the Beddit® and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (Mean difference ± 2SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. MD Mean Difference (or bias, in this panel a positive value indicates overestimation); LOA lower limits of agreement; PSG, polysomnography; TST, Total sleep time; SOL, sleep onset latency and SE sleep efficiency. Panel A total sleep time; Panel B sleep onset latency; Panel C sleep efficiency.

177x165mm (600 x 600 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4
	4	Study objectives and hypotheses	1, 5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	
Participants	**6**	Eligibility criteria	5-7
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	5-7
	8	Where and when potentially eligible participants were identified (setting, location and dates)	5-7
	9	Whether participants formed a consecutive, random or convenience series	5
Test methods	**10a**	Index test, in sufficient detail to allow replication	5
	10b	Reference standard, in sufficient detail to allow replication	6
	11	Rationale for choosing the reference standard (if alternatives exist)	7
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7-8
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	6
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	6
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	7-8
	15	How indeterminate index test or reference standard results were handled	7
	16	How missing data on the index test and reference standard were handled	10
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	5
	18	Intended sample size and how it was determined	-
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Figure1
	20	Baseline demographic and clinical characteristics of participants	Table 1
	21a	Distribution of severity of disease in those with the target condition	Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	-
	22	Time interval and any clinical interventions between index test and reference standard	-
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Table 2
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Table 3
	25	Any adverse events from performing the index test or the reference standard	-
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-12
	27	Implications for practice, including the intended use and clinical role of the index test	12
OTHER INFORMATION			
	28	Registration number and name of registry	-
	29	Where the full study protocol can be accessed	-
	30	Sources of funding and other support; role of funders	13

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.